



Universidad
Carlos III de Madrid

Departamento de Ingeniería Telemática

TRABAJO DE FIN DE GRADO

ANÁLISIS DEL COMPORTAMIENTO EN REDES SOCIALES DE PARTICIPANTES EN *MOOCS*

Autor: Javier Velo Beascoechea

Tutores: Carlos Alario Hoyos
Iria Estévez Ayres

Leganés, junio de 2016

Título: ANÁLISIS DEL COMPORTAMIENTO EN REDES SOCIALES DE PARTICIPANTES EN *MOOCs*

Autor: JAVIER VELO BEASCOECHEA

Director: CARLOS ALARIO HOYOS

EL TRIBUNAL

Presidente: _____

Vocal: _____

Secretario: _____

Realizado el acto de defensa y lectura del Trabajo de Fin de Grado el día ____ de _____ de 20__ en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO

PRESIDENTE

Agradecimientos

Agradezco a mi familia los ánimos que me ha dado durante toda la carrera y las *enhorabuenas* que me felicitaba cada vez que obtenía resultados positivos. En especial a mis padres también por su comprensión y por echarme una mano con algunas funcionalidades de *Word* y *Excel* durante la realización de este trabajo.

También a mis amigos por sus ánimos e incluso por ofrecerse a ayudarme con alguna cosa que necesitara.

Por supuesto, gracias a la Universidad por todos los conocimientos transmitidos durante estos cuatro años y por los valores que he adquirido durante mi experiencia universitaria. A todos los profesores que he tenido, ya hayan hecho las clases más o menos amenas, y a mis coordinadores de este trabajo por sus consejos y por guiarme en el desarrollo del mismo.

Y finalmente, gracias a todos los creadores de *MOOCs* por su trabajo y a todos los participantes de los mismos por su entusiasmo en esta *nueva* e interesante forma de enseñar y aprender.

Resumen

En este Trabajo de Fin de Grado se analizaron datos sobre un *Massive Open Online Course (MOOC)* sobre introducción a la programación en *Java*, llevado a cabo en la Universidad Carlos III de Madrid y desplegado en la plataforma *edX*. Dichos datos contenían las publicaciones de los usuarios en los foros de discusión, las notas finales de dichos usuarios y los mensajes intercambiados a través de la red social *Twitter* con el *hashtag* del curso (*#javaedxuc3m*).

A partir de los datos mencionados, este trabajo apuntaba a dos objetivos principales. Por un lado, comprobar si ciertas características de este *MOOC* coincidían con las de otros cursos estudiados en varios artículos de investigación, lo cual podrá facilitar la creación de modelos generales que ayuden a comprender e incluso contribuir a mejorar el desarrollo de futuros cursos. Por otro lado, investigar las cualidades de los *top contributors* e intentar identificarlos poco después de comenzar el curso. Éstos son los participantes que más aportaciones han realizado en un *MOOC* y poseen una especial importancia dado que en ocasiones pueden llegar a desempeñar acciones propias del equipo docente.

Mediante la programación de algoritmos para automatizar el proceso de obtención de información y generar distintas gráficas, se llegaron a varias conclusiones según los objetivos anteriormente mencionados. Por una parte, en cuanto a características generales del curso, se asemejan a las de otro *MOOC* que se tomó como referencia, como la gran afluencia de mensajes presentes los primeros días por la emoción del comienzo del curso. Por otra parte, se dan algunas conclusiones sobre los *top contributors* para este *MOOC*, como que responden más en hilos ya creados en vez de crear nuevas discusiones. Aunque también se da algún dato sobre su pronta caracterización, serán necesarios más estudios sobre este asunto, como por ejemplo examinar distintos *MOOCs* de temáticas y duraciones diferentes o buscar características comunes entre los perfiles de los mayores contribuidores.

Palabras clave:

Massive Open Online Courses (MOOCs), foro de discusión, redes sociales, *top contributors*.

Abstract

In this Bachelor's thesis, data about a Massive Open Online Course (MOOC) was analyzed. This course was an introduction to *Java* programming carried out in *Universidad Carlos III de Madrid* and deployed in the platform *edX*. Said data contained the users' contributions to the discussion forum, their final marks and the messages that were exchanged via the social network *Twitter* with the course *hashtag* (*#javaedxuc3m*).

From the aforementioned data, this work aimed for two main objectives. On one side, checking if certain characteristics from this MOOC coincided with the ones derived from other research papers that studied other courses. This could facilitate the creation of general models to help understanding these courses and contributing to improve the progress of future ones. On the other side, investigating top contributors' qualities and trying to identify them soon after the course starts. These are the users that make the most contributions in a course and they have a notable importance since in some cases they may perform actions expected from teachers.

By programming several algorithms to systematize the process of obtaining information and generating some graphs, different conclusions were noted according to the previous objectives. On one hand, respect to general course characteristics, they are similar to the ones from other MOOC used as reference, like the large number of messages during the first days of the course due to the excitement created around it. On the other hand, some conclusions about the top contributors of this course are given, like the fact that they mostly comment in already created threads instead of creating new ones. Although some data about their early identification are also given, additional studies of this issue will be required, as inspecting different MOOCs with distinct thematic and duration or searching for common characteristics among the top contributors' profiles, for example.

Keywords:

Massive Open Online Courses (MOOCs), discussion forum, social networks, top contributors.

Índice general

1. INTRODUCTION AND OBJECTIVES	1
1.1 Introduction	1
1.2 Objectives.....	3
1.3 Regulatory Framework.....	3
1.4 Methodology	4
1.5 Report structure	5
2. CONTEXTO Y TRABAJO RELACIONADO	8
2.1 Estado del arte sobre <i>MOOCs</i> y componente social	8
2.2 Tipos de <i>MOOCs</i>	10
2.3 <i>Flipped Classroom</i>	12
2.4 Tecnologías	13
3. CASO DE ESTUDIO	16
3.1 Problema principal	16
3.2 Características del curso.....	17
3.3 Foro y estructura de los mensajes.....	18
3.4 <i>Twitter</i>	20
4. DISEÑO E IMPLEMENTACIÓN DE LA SOLUCIÓN	23
4.1 Diseño de la solución	23
4.2 Programas diseñados	24
5. RESULTADOS	32
5.1 Curso	32
5.2 <i>Top contributors</i>	46
6. CONCLUSIONS AND FUTURE WORK.....	55
6.1 Conclusions	55
6.2 Acquired competences	58
6.3 Future work	59
REFERENCIAS.....	61

ANEXO I: PLANIFICACIÓN.....	65
ANEXO II: PRESUPUESTO.....	69
ANEXO III: INTRODUCCIÓN Y OBJETIVOS.....	72
III.1 Introducción.....	72
III.2 Objetivos.....	74
III.3 Marco regulador.....	74
III.4 Metodología.....	75
III.5 Estructura de la memoria.....	76
ANEXO IV: CONCLUSIONES Y LÍNEAS FUTURAS.....	79
IV.1 Conclusiones.....	79
IV.2 Competencias adquiridas.....	82
IV.3 Líneas futuras.....	83
ANNEX V: SUMMARY.....	86
V.1 Introduction and objectives.....	86
V.2 Context and related work.....	88
V.3 Case of study.....	90
V.4 Solution design and implementation.....	91
V.5 Results.....	104
V.6 Conclusions and future work.....	109

Índice de figuras

Figura 1. Comparación entre el modelo tradicional y el modelo <i>flipped</i> [21].	13
Figura 2. Inicios de discusión y comentarios en discusiones ya creadas.	20
Figura 3. Porcentaje de aprobados y suspensos.	33
Figura 4. Porcentaje de aprobados que comentaron y no comentaron en el foro de discusión.	33
Figura 5. Porcentaje de suspensos que comentaron y no comentaron en el foro de discusión.	34
Figura 6. Notas obtenidas por los participantes excluyendo a los que sacaron un cero. Nota: no se han excluido a los coordinadores y <i>staff</i>	35
Figura 7. Número de comentarios por fechas.	36
Figura 8. Histogramas representando el número de usuarios que realizaron diferentes cantidades de aportaciones al foro de discusión.	37
Figura 9. Histograma dividido en intervalos con diferentes longitudes para una mejor visualización (todo el curso). Nota: no se han excluido a los coordinadores y <i>staff</i>	38
Figura 10. Histograma dividido en intervalos con diferentes longitudes para una mejor visualización (primeras dos semanas de curso). Nota: no se han excluido a los coordinadores y <i>staff</i>	39
Figura 11. Número de mensajes en cada franja horaria (todo el curso; formato UTC). Nota: el valor que corresponde a una hora, X, es el número de mensajes publicados entre X:00:00 y X:59:59.	40
Figura 12. Número de mensajes en cada franja horaria (eliminando los 3 primeros días; formato UTC). Nota: el valor que corresponde a una hora, X, es el número de mensajes publicados entre X:00:00 y X:59:59.	41
Figura 13. Para todo el curso, porcentaje de alumnos que, con al menos un número dado de aportaciones al foro de discusión, aprobaron el <i>MOOC</i>	42
Figura 14. Para las dos primeras semanas del curso, porcentaje de alumnos que, con al menos un número dado de aportaciones al foro de discusión, aprobaron el <i>MOOC</i>	43
Figura 15. Fechas de publicación de la última aportación al foro de los alumnos que suspendieron.	44

Figura 16. Fechas de publicación de la última aportación al foro de los alumnos que aprobaron.....	45
Figura 17. Porcentaje de alumnos aprobados que, para una fecha dada, siguen realizando comentarios.	45
Figura 18. Evolución temporal de los comentarios desde el mayor contribuidor número 1 hasta el mayor contribuidor número 5.	46
Figura 19. Evolución temporal de los comentarios desde el mayor contribuidor número 6 hasta el mayor contribuidor número 10.	47
Figura 20. Evolución temporal de los comentarios desde el mayor contribuidor número 11 hasta el mayor contribuidor número 15.	47
Figura 21. Evolución temporal de los comentarios desde el mayor contribuidor número 16 hasta el mayor contribuidor número 20.	48
Figura 22. Diagrama social de todos los alumnos del curso.	51
Figura 23. Diagrama social considerando como “respondedores” a los 20 <i>top contributors</i>	52
Figura 24. Diagrama social considerando como “respondedores” a los 10 <i>top contributors</i>	53
Figura 25. Esquema de la planificación del proyecto.	66
Figura 26. Diagrama de <i>Gantt</i> del proceso de organización del proyecto. Nota: no tener en cuenta las horas; fijarse solamente en los días y entenderlos como días completos. .	66
Figure 27. For the whole course, percentage of students that, with at least a given number of contributions to the forum, passed the MOOC.	94
Figure 28. For the first two weeks of the course, percentage of students that, with at least a given number of contributions to the forum, passed the course.....	94
Figure 29. Histograms representing the number of users that contributed with different number of messages to the forum.	95
Figure 30. Histogram divided in intervals with different lengths for a better interpretation (the whole course). Note: moderators and staff were not excluded.	95
Figure 31. Histogram divided in intervals with different lengths for a better interpretation (first two weeks of the course). Note: moderators and staff were not excluded.	96
Figure 32. Number of messages in each time period (the whole course; UTC format). Note: the value corresponding to a time, X, is the number of contributions between X:00:00 and X:59:59.	96
Figure 33. Number of messages in each time period (excluding the first 3 days; UTC format). Note: the value corresponding to a time, X, is the number of contributions between X:00:00 and X:59:59.	97
Figure 34. Social diagram considering as “repliers” the top 20 contributors.	98
Figure 35. Social diagram considering as “repliers” the top 10 contributors.	99
Figure 36. Dates of the last contribution to the forum of the students that failed the course.	100
Figure 37. Dates of the last contribution to the forum of the students that passed the course.	100
Figure 38. Percentage of students that passed that, for a given date, keep on commenting .	101
Figure 39. Temporal evolution on the number of comments from top contributor 1 to top contributor 5.	101
Figure 40. Temporal evolution on the number of comments from top contributor 6 to top contributor 10.	102
Figure 41. Temporal evolution on the number of comments from top contributor 11 to top contributor 15.	102

ÍNDICE DE FIGURAS

Figure 42. Temporal evolution on the number of comments from top contributor 16 to top contributor 20.	103
Figure 43. Marks obtained by the students excluding zeros. Note: moderators and staff have not been excluded.	103
Figure 44. Number of comments each date.....	104
Figure 45. Percentage of passing and failing students.	106
Figure 46. Percentage of passing students that contributed and did not contribute to the forum.	106
Figure 47. Percentage of failing students that contributed and did not contribute to the forum.	107

Índice de tablas

Tabla 1. Síntesis comparativa de diferentes tipos de <i>MOOCs</i> [14].	11
Tabla 2. Estructura de los mensajes del foro.....	20
Tabla 3. Clasificación de los <i>tweets</i>	21
Tabla 4. <i>Scripts</i> de <i>MATLAB</i> creados	25
Tabla 5. Número de aportaciones al foro y número y porcentaje de días aportados de los 20 mayores contribuidores. Nota: el curso duró 64 días (del 28 de abril al 30 de junio de 2015).....	48
Tabla 6. Número de aportaciones al foro y número y porcentaje de días aportados de los 20 mayores contribuidores durante las dos primeras semanas del curso.	49
Tabla 7. Número y porcentaje de inicios de discusión y respuestas a discusiones ya creadas de los 10 mayores contribuidores.....	50
Tabla 8. Horas dedicadas a este Trabajo de Fin de Grado.	67
Tabla 9. Estimación del presupuesto.....	70
Table 10. Scripts designed with <i>MATLAB</i>	93
Table 11. Classification of the <i>tweets</i>	105
Table 12. Number of contributions to the forum and number and percentage of contributed days of the top 20 contributors. Note: the MOOC lasted for 64 days (from April 28 th to June 30 th of 2015).	108
Table 13. Number of contributions to the forum and number and percentage of contributed days of the top 20 contributors during the first two weeks of the MOOC.	109
Table 14. Number and percentage of started discussions and comments in already created discussions of the top 10 contributors.	112

Chapter 1¹

Introduction and objectives

1.1 Introduction

Massive Open Online Courses (MOOCs due to their abbreviation in English; *Cursos Masivos en Línea y Abiertos* in Spanish) have captured the interest of many high-level education institutions. Their defenders argue that MOOCs make access to education easier to as many people as possible, increase the visibility of the institution which creates them and allow teachers to experiment with online education to a wide and diverse number of students [1]. On the other side, people who oppose to MOOCs criticize them as virtually harming, as technology disruptive of education that offers a watered-down formal learning and as a risk of even more cuts in government-funded school budgets [2]. Even though this document will not participate in this debate, it should be accepted that MOOCs have considerably impacted on the way of teaching and learning (although they will probably not replace traditional face-to-face teaching, at least for now). Actually, in one of the referenced documents, its authors asked themselves whether MOOCs would replace university education or not [3]. In that document, MOOCs are considered not as an alternative to post-secondary education but as its complement. Besides, an example is given comparing a computer and a tablet: one is not a substitute of the other; both accomplish different purposes. Thus, the same way MOOCs have features that traditional teaching methods do not, face-to-face education has characteristics which MOOCs lack.

¹ Para la traducción en español de este capítulo, consultar el Anexo III.

Trying to explain **typical features** of MOOCs [4] [5], they are Internet-offered courses that rely on different platforms like *MiríadaX* or *edX*. A list of offered courses and information about them is typically available to anyone surfing the Web, even though a registration is required in order to enroll on the course. When selecting a MOOC, this will usually show additional information like its duration and start date, among others. In addition, sometimes it can also be found a description of the course and different user ratings if it is not the first time the MOOC is taught, and even a video introducing the course. Once registered on the course and once it starts, different material provided by the organizers will be given, like explicative videos, documents dealing with the concepts and tasks for the student to be evaluated. This student will be able to complete these tasks as he or she wishes as long as it is done before the assignment deadline. Said assignments can be split in two groups. On one side, watching videos and others contents including information about the subject. On the other side there are exercises, which can be divided in two subgroups. On one hand, close-ended questions like multiple-choice tests in which there exists an objective and predefined answer. These can be passed through an automatic system that corrects them. On the other hand, open questions in which there is not a predefined answer and there exists certain degree of subjectivity. In MOOCs that contain these types of questions, like the one studied in this document, it is usually used a technique called “peer review”, in which a student grades the work done by other users. In the end, each participant receives a final mark according to his or her results throughout the different course activities. Normally, taking the course is free of charge but obtaining a certificate may incur in a monetary expense. Actually, from 2014, the three most famous platforms (*Coursera*, *edX* and *Udacity*) demand payment to users who wish to obtain a certificate [6]. On average, these certificates cost around \$50 (roughly €45).

The field that comprehends MOOCs is considerably new, which implies a lack in studies and research papers addressing them. Regarding the newness of these courses, 2012 was claimed to be “the year of the MOOC” [7] since the most famous platforms aforementioned were founded in this year: *Coursera*, *edX* and *Udacity*. Regarding Spanish courses, the platform *MiríadaX* is also well-known, although it also has Portuguese and English courses. It was founded in 2012, too.

MOOCs have **social tools** (discussion forums and accounts in different social networks, among others) in which the students interact with each others by formulating questions or answering them, sharing possibly interesting links or giving feedback to the teachers regarding the material, amongst other actions. Along these lines, and in the same way in a traditional face-to-face course there may exist students significantly active (by participating more than their classmates or volunteering for different activities, for instance), inside the MOOCs discussion forums something similar may happen. There exists the possibility that some students will become more involved in the forums than their course mates, by checking and commenting about already-posted answers or by reporting errors to the staff more than their peers, among other possibilities mentioned above. In this way, participants that make the most contributions will become the course “top contributors”. As it will be seen later, it may be interesting to identify these contributors early. Bringing forward some of the reasons, as top contributors often carry out actions attributable to teachers, they may result in great help towards other participants. They can also assist in keeping a stable and pleasant environment around the discussion forums [1].

1.2 Objectives

This work studies a MOOC organized by *Universidad Carlos III de Madrid*, which took place between April and June of 2015, using the platform *edX*. The MOOC title was “**Introduction to Programming with Java Part 1: Starting to Program with Java**”. Based on the data provided by said platform regarding student’s final marks and their contributions to the forum and *Twitter*, two objectives can be suggested.

On one side, it can be checked whether the **hypotheses proposed by other scientific papers** addressing other MOOCs are fulfilled or not for this case of study. These hypotheses and other conclusions will be dealt with in Chapter 2. Anticipating slightly these issues, a relevant hypothesis could be the one addressing the correlation between the users that contribute the most and the ones that obtain the best marks, and a notable conclusion could be the one that states that it should be taken into account not only the number of contributions but the quality of these, too. The usefulness of these results could allow the creation of models that can be generalized to all MOOCs, in order to help understand these courses and contributing to improve the progress of future ones. As examples it could be specified understanding how students are connected with each others via the discussion forums, predicting which percentage of students will finalize the courses, understanding the causes by which a participant continues or drops out the MOOC or discovering whether the top contributors answer to any other student or just to a small group of participants.

On the other side, from what was mentioned before it can be deduced that these top contributors play an important role in these courses since they sometimes perform tasks that are expected from teachers and staff. This way, it could be interesting to **identify these top contributors early** in the course and even grant them some kind of moderator benefit. Certainly, this is a prediction problem that would be interesting to solve as soon as possible after the start of the MOOC. As it is specified in Chapter 2, there exist some impediments that hinder this prediction, like the large number of messages exchanged in the beginning on the course.

1.3 Regulatory Framework

During the execution of this project, access to three files was granted: one of them had the “tweets” exchanged with the “hashtag” of the course (*#javaedxuc3m*) via the social network *Twitter*, the second one contained the messages of the discussion forum, and the last one had the marks obtained by the participants. The three files had different fields apart from the text of the message itself, like the authors’ username (in Chapter 3 a table with the fields of the forum messages will be seen). The third file also contained the real names of the participants. Due to this, it is needed to talk about the law that regulates the management of personal information; that is, to talk about the *Ley Orgánica de Protección de Datos* [8].

The just-mentioned Law has as purpose to protect and guarantee the rights of the people whose personal information is handled. The regulation established by this Law prevents the use of personal data to invade people's privacy and violate their public liberties. A statement that could be remarked is the one that says that management of personal information must be in compliance with the purpose for which it was collected. It could also be noted that people whose personal information is being handled must have given permission to use such data. This will be related later with *edX* privacy policy and terms of use.

As stated in the *Guía de Seguridad de Datos* [9], a three-level classification of personal data can be established. According to this classification, different security measures will be required regarding the use of this data. These levels are high, medium and basic. After reading the properties of each level declared in the guide, it is concluded that the information used in this project belongs to the basic level.

As mentioned before, in addition to the *Ley de Protección de Datos*, it also should be taken into account what is stated in the website of the platform *edX* itself. On one side, in relation to the **terms of use** [10], it could be referenced the section "user postings". In this one it is stipulated that, once a user posts a comment, he or she is granting an irrevocable license to display, transfer and distribute (among other actions) this comment. In the same way, all *edX* website users can access this message for their own purposes. On the other side, regarding its **privacy policy** [11], in section "usernames and postings" it is said that the comments and other information published by the users may be viewed by people who visit the website, so it is advised not to share any data that can be used to identify the person behind the username. Besides, information may be shared with third parties, among other reasons, for scientific researches, especially the ones treating about education. This is the one that applies in this work since it is a Bachelor's thesis with a research component about MOOCs and education based on a course of this platform. Once again, even though access to the real usernames was granted, the personal information utilized was usernames, identifiers and obtained mark.

For all these reasons, it is justified the collection and use of personal data utilized during the development of this work respecting the current regulatory framework.

1.4 Methodology

Firstly, the person in charge of this work, after studying the list and description of each Bachelor's thesis noted in the university webpage, had a meeting with this project academic coordinators. In this way, he started by reading some papers provided by these coordinators. In the same way, the student searched for and read some additional documents. The papers given by the teachers dealt with another MOOC used as case of study (called "Digital Education of the Future") and they noted different conclusions from their investigation. The purpose of such readings was making notes about which data could be used as starting point and which lines could be studied. It could be remarked that the student already had some knowledge regarding MOOCs since in his third year of degree he coursed a subject that was about them (the subject title was

Conocimiento Libre y Aprendizaje en la Web; a translation would be “Free Knowledge and Learning on the Web”).

After all this, and once the student decided to stay with this project, the academic coordinators granted him a file containing the course “tweets” already mentioned before. Based on this file, a table containing the number of messages written by each user was created. Next, after another meeting to talk about the project and the aforementioned classification, another file was given to the student, this one containing the forum messages of the MOOC which was being studied. From that point on, the followed methodology consisted of scheduled meetings in which the responsible of the project showed the progress achieved to the coordinators. In said meetings new lines of research were proposed and different approaches to the obtained results were discussed. This is, the process can be classified as recurring (communication with the coordinators in order to validate the obtained results and propose new ideas) and incremental (it starts with simple analysis about the available data and it continues towards more complex graphs or algorithms).

Finally, during the fulfillment of this report, by means of email, the student shared his progress with the coordinators and these suggested improvement for said report. Before this, it was discussed which report structure could be suitable, and that structure will be developed in the next section. I can also be noted that the writing of this work has been a continuous and iterative process from the beginning of the project.

1.5 Report structure

This report begins with an introductory chapter. In it, an introduction to the work is presented and the related context is mentioned. Once this issue is dealt with, the primary objectives are presented, followed by a reference to the regulatory framework relating management of personal information provided by *edX* and a brief commentary about the followed methodology in the development of this project.

After this chapter, in the second one, the related context is treated more deeply than in the introduction, at the same time some useful conclusions towards this work are remarked. Also, the technologies used in this project are noted; some were already known (like *MATLAB*) and some others were required to learn (like *Gephi*).

Then, in the third chapter, characteristics of the case of study of this work are specified, being this a MOOC carried out by *Universidad Carlos III de Madrid* between April and June of 2015. In this way, it is also explained the challenges and problems this project faced. The design of the MOOC and the characteristics of the messages of the course forum contained in the file provided by are also commented.

In the fourth chapter, the procedures carried out to solve the problems stated in chapter three and the actions taken to reach the proposed objectives are explained. A table is presented containing the names of the scripts designed with *MATLAB* as well as a brief description of each one. Additionally, after the table, these algorithms are explained by using pseudocode.

CHAPTER 1: INTRODUCTION AND OBJECTIVES

Then, in the fifth chapter, the obtained results are showed and briefly discussed. Firstly, some general regarding the course are given. Then, several results with respect to the top contributors are discussed.

In the last chapter, the conclusions based on the results of chapter five can be found. Apart from this, it is also stated which abilities the student has developed during this work and different lines relating future projects.

As appendixes, there is a planning summarizing the development of this project and an estimated budget. A Spanish version of the introduction and conclusion can also be found. Finally, there is an English summary of the whole work.

Capítulo 2

Contexto y trabajo relacionado

2.1 Estado del arte sobre *MOOCs* y componente social

Como parte del proyecto, al comienzo de éste se leyeron varios artículos recientes sobre estudios similares en *MOOCs*. En ellos se hacían algunas clasificaciones de los alumnos participantes y se incluían varios consejos con vistas a otros proyectos que fueran a abarcar el ámbito de estos cursos. Se recogerán en esta sección las conclusiones más notables que se obtuvieron de dichos textos.

Primero, en los *MOOCs* se puede destacar una **diferencia** considerable respecto a la impartición de clases tradicionales “cara a cara”, siendo ésta la cantidad de alumnos matriculados (también existen cursos *online* “cerrados” que comparten el número limitado de estudiantes con los cursos cara a cara). Por una parte, en estos cursos tradicionales, el número de alumnos en general es menor (pocas decenas de estudiantes) y se les presupone dispuestos a afrontar tales cursos y a participar activamente en ellos, debido en gran parte al coste económico que generalmente supone la matriculación. En cambio, en los *MOOCs*, dado que la matriculación suele ser gratuita, el número de participantes aumenta a cientos e incluso miles (aunque pueden existir costes alternativos como, por ejemplo, el correspondiente a la obtención de un certificado). Aún así, este aspecto tiene un matiz, tratándose éste de que el número de alumnos que acaba un *MOOC* suele ser inferior al 10% de los que se matriculan en él [12]. Aunque este porcentaje parezca pequeño, la cantidad en sí suele seguir siendo del orden de cientos de alumnos,

por lo que generalmente sigue estando por encima del número de aprobados en cursos impartidos de manera tradicional.

En relación con el párrafo anterior y centrando la atención de nuevo en los *MOOCs*, los alumnos que tengan más interés en el curso serán, habitualmente, los que realicen mayores aportaciones a las herramientas colaborativas, siendo el foro la más utilizada en general. Por una parte, esto es de ayuda a la hora de identificar a los *top contributors* (se recuerda que, con este término, se designa a los usuarios más activos en las herramientas colaborativas del curso). Pero, por otra parte, hay que tener en cuenta la calidad de estos mensajes, ya que un estudiante puede haber realizado muchas aportaciones al foro y no haber contribuido con nada realmente útil. Además, hay un **problema** adicional si se quiere identificar de manera temprana a los mencionados *top contributors*, siendo éste la concentración de mensajes en el foro al comienzo del curso por la emoción (*hype* o *excitement*) que genera, sobre todo si se abre una discusión de bienvenida en el mismo [13] [1]. Así, hay mucha presencia de ruido a la hora de seleccionar pronto a los contribuidores *top*. Otro problema para la identificación de estos mayores contribuidores es la existencia de usuarios que se registran en el curso cuando éste ya ha comenzado (apodados como *latecomers* en algunos documentos [13]). En alguna ocasión [13], un porcentaje interesante de los alumnos que aprueban el curso pertenecen a este grupo (el 26,1% de los que aprobaron el curso “Educación Digital del Futuro” eran *latecomers*; si se pasa este porcentaje a números concretos, se obtienen 119 *latecomers* de 456 aprobados [13]). Este hecho puede dificultar un poco este tipo de investigaciones dado que probablemente se tengan pocas contribuciones de este grupo de alumnos en este periodo inicial del curso (los *latecomers* se caracterizan, efectivamente, por no estar presentes durante el inicio del curso) y puede que alguno de ellos llegue a convertirse en *top contributor*. El curso mencionado anteriormente, “Educación Digital del Futuro”, tiene una especial relevancia en este estudio, puesto que es el que usan como caso de estudio concreto dos de los documentos utilizados como referencias [13] [1].

También se pueden encontrar algunos resultados que específicamente están expuestos como **punto de partida** para proyectos como éste. Dichos resultados fueron conclusiones del estudio de otro *MOOC*. Tres de ellos son no tener en cuenta el género del alumno, realizar las predicciones cuando haya tenido lugar la primera actividad de evaluación y tener en cuenta las contribuciones colectivas a todas las herramientas por parte de cada alumno en vez de hacer grupos distintos [1]. Además, se aconseja no tener en cuenta solamente el número de contribuciones sino añadir factores como la calidad de las mismas o en qué periodo fueron publicadas. Otra opinión que aparece en varios documentos [13] [1] es la de aplicar el concepto de *gamification* (en español conocido como “ludificación”, “gamificación” o “juguetización”). De lo que se trata es de utilizar técnicas y dinámicas propias de los videojuegos para crear un ambiente más divertido y llevadero en lo que *a priori* puede parecer una actividad aburrida. Estas técnicas pueden ir desde la concesión a los estudiantes de algún tipo de medalla o insignia hasta permitirles modificar con características especiales sus nombres de usuario o imágenes de perfil. Este aspecto de *gamification* influye en la motivación de los alumnos: si éstos están más motivados, probablemente estarán más dispuestos a participar más activamente en el curso, tanto completando las actividades del mismo como realizando contribuciones a los foros. Se trata, efectivamente, de pequeños detalles que pueden ayudar a la formación de una mejor y más agradable experiencia por parte de los participantes.

Finalizando, se pueden mencionar algunos apuntes misceláneos. Por ejemplo, se puede pensar que adquirir la certificación de que se ha aprobado el curso no significa necesariamente que el alumno haya aprendido mucho [12] (aunque esta situación debería ser improbable a menos que el curso esté mal diseñado o que el participante haga trampas en las evaluaciones). Esto es, podría darse el caso de un *latecomer* que no ha podido recibir certificación (por incorporarse de forma tardía al curso) y que ha adquirido muchos conocimientos, mientras que alguien que aprueba el curso y obtiene un certificado puede no haber aprendido demasiado. Otro apunte destacado es el de disponer de varias herramientas sociales en los *MOOCs* para que cada alumno utilice aquella con la que se sienta más a gusto, además de obtener diferentes resultados y niveles de participación (aunque el foro del curso suele ser la herramienta más utilizada) [13]. Y para terminar, recordar alguna utilidad que puede tener el identificar a los contribuidores *top* de manera temprana. En plataformas como *edX*, los profesores pueden asignar ciertos privilegios a estos alumnos, como permitir que editen o eliminen mensajes escritos por sus compañeros [1]. De esta manera, se forja una colaboración para mantener adecuadamente el foro y se podría mejorar el apoyo dado a los participantes de los *MOOCs* promoviendo discusiones en torno a estos cursos (esto es, el hecho de que el foro sea activo, agradable y eficiente fomenta el que más usuarios participen en él).

2.2 Tipos de *MOOCs*

En poco tiempo desde el surgimiento de los *MOOCs* se ha producido tal evolución y difusión de los mismos que, aun guardando rasgos comunes, no responden a un modelo único [14]. De este modo, existen **varios enfoques según la teoría pedagógica** a la que más se asemejan [14] [15].

El primero de estos enfoques engloba a los denominados *cMOOCs*, en los cuales la “c” inicial proviene de la **teoría conectista o conectivista** [16]. Dicha teoría pronuncia que el aprendizaje es un proceso asentado en la capacidad de construir, unir y mantener fuentes de información especializadas enlazadas por medio de redes. De esta manera, los *cMOOCs* están orientados a la interacción entre los usuarios para generar conjuntamente conocimiento. Así, en esta especie de aprendizaje colaborativo, el elemento principal es el propio alumno (y no el profesor) quien, interactuando con los recursos disponibles en plataformas abiertas en red, crea y co-genera nuevo contenido. Así pues, los estudiantes no trabajan en grupos, sino en red, e interactuando como colectivo en herramientas tales como blogs, redes sociales o foros. Por todo ello se ve que los *cMOOCs* se alejan de la forma de enseñar tradicional dado que enfatizan en la capacidad de los alumnos para interactuar, organizarse ellos mismos y generar conocimientos de forma conjunta, cooperando y trabajando en red. En contrapartida, todas estas características limitan el ámbito de difusión de los *cMOOCs* puesto que resultan difíciles las tareas de evaluación mediante pruebas objetivas y de generación de certificados.

El segundo modelo a citar engloba a los *xMOOCs*, en cuales la “x” inicial viene de plataformas como *edX*, *MiríadaX* o *MITx* [17] [18]. Estos *xMOOCs* se basan en una pedagogía más tradicional, apoyándose en la **teoría conductista** del aprendizaje [19]. Ésta mantiene al profesor como protagonista de todo el proceso de aprendizaje mediante una enseñanza explícita y directa. Dicho profesor transmite los conocimientos, diseña los

objetivos y evalúa los resultados. En este modelo, aunque también pueden existir foros de discusión e intercambios de ideas entre los alumnos, la estructura resulta más rígida que en los *cMOOCs* dado que esta interacción ocurre en torno a una misma página web en la que el profesor tiene el control. Continuando con las características de los *xMOOCs*, éstos suelen facilitar módulos con material textual o vídeos, entre otros, todos ellos dirigidos hacia la obtención de un cierto conocimiento. De esta forma, resultan más sencillas las acciones de evaluación y distribución de certificados ya que es posible utilizar métodos tradicionales como cuestionarios y entrega de trabajos. Por todo ello, y sobre todo por los últimos aspectos mencionados, este modelo es el adoptado por muchas universidades. El curso que estudia este proyecto es, en gran medida, un *xMOOC*.

Cabe destacar que esta clasificación en modelos va quedando obsoleta puesto que, de manera continua, surgen cursos que presentan alguna peculiaridad respecto a los dos modelos anteriormente descritos. Por ello, entre los prototipos propuestos, la bibliografía consultada [14] [15] destaca un tercer modelo denominado *tMOOCs*, en el cual la “t” inicial viene de la **orientación a tareas** de los mismos. En este modelo se destaca la importancia de cooperación y comunicación entre los estudiantes pero siempre como apoyo hacia la realización de tareas, que es el objetivo principal. De la misma forma, el contenido se presenta distribuido en diferentes formatos pero, de nuevo, siempre con el objetivo de realizar ciertas actividades para completar la adquisición de conocimientos. Así, se concluye que los *tMOOC* constituyen un modelo híbrido entre las teorías conectivista y conductista.

Finalmente, se ve que cada modelo posee unas características propias y cada uno se orienta a la persecución de unos objetivos propios. A continuación, se aporta una tabla (Tabla 1) en la que se comparan los distintos modelos aquí descritos.

	<i>cMOOCs</i>	<i>xMOOCs</i>	<i>tMOOCs</i>
Teoría de aprendizaje	Conectivismo	Conductismo	Híbrida
Modelo de aprendizaje	Co-generación de conocimiento	Adquisición (clásica) de conocimiento	Adquisición y generación de conocimiento
Orientado a...	Redes	Contenidos	Tareas
Conocimiento	Distribuido y co-generado	Declarado por la organización del curso	Declarado por la organización del curso
Contenido	Co-creación o socio-creación a partir de un tema central	Tema central; desarrollo enfocado a conceptos	Tema central; desarrollo centrado a actividades
Objetivos del aprendizaje	Autodefinido por cada participante	Definido por el profesor	Definido por el profesor
Ámbito del aprendizaje	Tema central abierto; interdisciplinar	Tema central cerrado; disciplinar	Tema central cerrado; disciplinar
Interacción	Fuerte	Débil	Media
Cohesión y control del curso	Participantes	Profesor	Profesor
Evaluación	Muy difícil o inexistente	Sencilla; automatizada en múltiples casos	Difícil o muy difícil
Preponderancia de...	Exploración y comunidad sobre los contenidos	Adquisición de contenidos sobre la comunidad	Tareas sobre la comunidad

Tabla 1. Síntesis comparativa de diferentes tipos de MOOCs [14].

2.3 *Flipped Classroom*

En esta sección se va a tratar brevemente sobre un modelo usado en cursos tradicionales “cara a cara” que también utiliza la enseñanza en línea y que es considerablemente novedoso. En este sentido, se puede hablar de la **evolución de los MOOCs hacia su uso en el campus**. Hay más metodologías, pero el modelo *flipped classroom* es de los más comúnmente implementados y el que se tratará a continuación.

Aunque para entonces aún no se había afianzado el término, en la década de 1990 aparecieron documentos cuyo contenido apuntaba hacia la metodología *Flipped Classroom* [20] (su traducción sería algo como “Clase Invertida”). Se trata de un modelo de trabajo con el que están experimentando varios docentes [21] [22]. Actualmente, en general, sobre todo en los niveles superiores de enseñanza, el tiempo de clase es dedicado a explicar la materia y transmitir a los alumnos las ideas básicas de cada unidad por parte del profesorado (clase magistral), mientras que el tiempo en casa se dedica a la realización de tareas. Entonces, lo que sugiere el modelo *Flipped Classroom* es, efectivamente, lo contrario: trasladar los procesos de aprendizaje fuera del aula y que el tiempo en clase sea utilizado para el desarrollo de tareas prácticas y potenciar la adquisición de conocimientos.

Se explica ahora con más detalle **en qué consiste este modelo**. Si se sigue esta metodología, el material docente facilitará al alumnado materiales audiovisuales de corta duración en los que se presentará la unidad y los conceptos principales. Dicho material podrá ser de creación propia o podrá ser un recurso ya existente en la red. Ahora bien, este modelo es mucho más que la edición y publicación de un vídeo. Se trata de un enfoque integral que combina el aumento de implicación e interés de los alumnos con la materia del curso al mismo tiempo que enlaza la instrucción directa con técnicas constructivistas. Este aspecto es precisamente la relación que guarda esta metodología con los MOOCs. En la Figura 1 está representada una comparación entre el modelo *flipped* y el modelo tradicional [21].

Como **ventajas** de este método, la más inmediata es el hecho de que los profesores envían conocimiento a casa en lugar de enviar problemas. De esta forma, el tiempo de clase se dedica a trabajar en las tareas de cada temario con la ayuda del profesor y del resto de alumnos. El tiempo en clase queda así abierto a discusiones, preguntas y actividades aplicadas que fomenten la articulación, exploración y aplicación de las ideas. Además, este método permite a los docentes dedicar una mayor cantidad de tiempo para ver las necesidades individuales de cada uno de sus alumnos. Por otro lado, la **desventaja** más notable que cabe pensar es la situación que se da cuando varios alumnos no llevan a cabo las tareas que deberían haber realizado previamente a la clase. Otra desventaja que se puede destacar es la dificultad añadida de introducir esta metodología si algún alumno no dispone de los medios necesarios para conectarse a Internet en su casa.

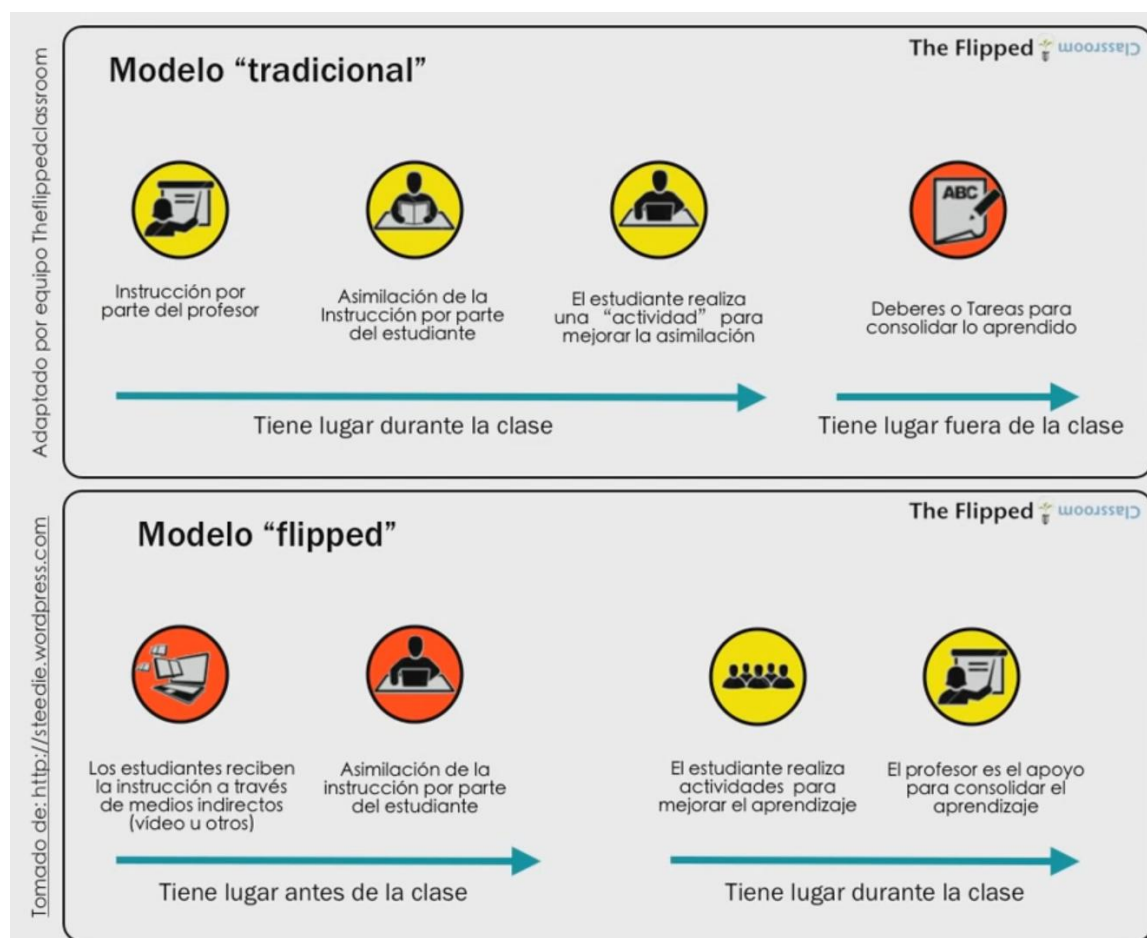


Figura 1. Comparación entre el modelo tradicional y el modelo *flipped* [21].

2.4 Tecnologías

Respecto a las tecnologías y herramientas utilizadas para la realización de este proyecto, la primera que cabe mencionar es *Notepad++*. Gracias a este programa se pudo visualizar correctamente el contenido del archivo con los mensajes del foro de discusión para examinar el formato de éstos y, a partir de entonces, buscar formas de tratar con ellos.

En cuanto a este archivo, y destacando la segunda tecnología, su extensión era *.mongo*, el cual fue generado por *MongoDB*, una base de datos basada en *NoSQL* [23]. En este tipo de bases de datos, aunque se soportan consultas con lenguaje *SQL* (*Structured Query Language*), la información no necesita estar almacenada en estructuras fijas como tablas [24].

La tercera tecnología que cabe destacar es *Microsoft Excel*, un *software* basado en hojas de cálculo [25]. Esta herramienta se utilizó para la obtención de algunas de las gráficas. Con ella también se abrió el documento facilitado por *edX* mencionado en el párrafo anterior y, tras realizar diferentes ajustes (ya que los datos no se leían de forma óptima), se pudo utilizar *Excel* como herramienta de la que extraer los datos que serían

CAPÍTULO 2: CONTEXTO Y TRABAJO RELACIONADO

utilizados en diferentes códigos de *MATLAB*. Por último, esta herramienta también sirvió para la ordenación y el almacenamiento de la información que se iba obteniendo.

Como cuarto programa al que mencionar está *MATLAB*, ya nombrado antes. Se trata de una herramienta de cálculo para diseñar algoritmos y visualizar gráficos [26], entre otras funcionalidades. En este programa se escribieron algoritmos para automatizar los procesos de obtención de información a partir de los datos de los que se disponían. También se utilizó para la representación de algunas gráficas.

Además de lo anterior, cabe destacar dos programas que se probaron para realizar diagramas sociales. El primero de ellos es *NodeXL* [27], una plantilla de acceso libre implementada en *Excel* que permite realizar este tipo de diagramas. Como segunda opción, también se probó a utilizar *Gephi* [28], un *software* gratuito para la visualización y exploración de gráficas y redes. Tras probar ambos, las gráficas presentes en este documento han sido obtenidas utilizando *Gephi*.

Para la escritura de esta memoria se utilizó *Microsoft Word*, herramienta de creación y edición de textos [29]. Este programa también fue utilizado para abrir y trabajar con el archivo de los *tweets* facilitado por *edX*. Además, solamente como mención, se usó *ProjectLibre* para la generación del diagrama de *Gantt* [30]. Éste es un programa de acceso libre para la planificación de proyectos.

Por último, se puede citar brevemente la propia plataforma *edX* y la distribución del contenido y de los foros en la misma. En cuanto a los contenidos, están divididos por semanas, y cada semana posee diferentes subapartados en los que puede haber un vídeo, unos ejercicios de autoevaluación, un texto explicativo, un examen o una actividad de revisión entre pares. Por parte de los foros de discusión, pueden ser filtrados de acuerdo a las diferentes subsecciones de cada semana. Dentro de estas subsecciones aparecen los hilos creados (del tipo “pregunta” o “discusión”) y el número de comentarios ya presentes en los mismos. Se aportan más detalles en el Capítulo 3.

Capítulo 3

Caso de estudio

3.1 Problema principal

En el capítulo introductorio ya se presentaron los objetivos que persigue esta investigación. Por una parte, se propone estudiar las características del curso en torno al foro de discusión y comprobar los resultados con conclusiones sacadas de otros artículos analizados en el Capítulo 2. Por otra parte, también está la investigación de la forma de actuar de los denominados *top contributors*.

De esta manera, se tiene como problema principal la obtención de información (valores numéricos concretos, porcentajes, gráficas, etc.) a partir de los datos de los que se disponía al comienzo de este proyecto. Una vez obtenida dicha información, se interpretará la misma y se obtendrán conclusiones. Ahora bien, cabe destacar que, dado que se realizaron **13778 contribuciones** al foro, la forma factible de realizar un estudio como éste es mediante la **automatización** de la lectura de datos (utilizando algoritmos y programas desarrollados en el ámbito del proyecto) y de presentación de resultados. De este último asunto se tratará posteriormente. Concretamente, en el Capítulo 4 (Tabla 4) se presenta una tabla en la que quedan apuntados los algoritmos programados en *MATLAB*, al igual que una breve descripción de los mismos junto con su tiempo medio de ejecución y número de líneas de código.

3.2 Características del curso

Recordando y ampliando lo que se mencionó con anterioridad, este trabajo se centra en torno a un *MOOC* transcurrido entre el 28 de abril y el 30 de junio de 2015 titulado *Introduction to Programming with Java Part 1: Starting to Program with Java* publicado en la web de *edX*.

Como información general del curso que se podía leer en su propia página (no se referencia con una dirección web porque, una vez que se acaba el curso, esta página desaparece; en estos momentos, se puede referenciar la versión de 2016 del curso [31]), se trata de una introducción a la programación en el lenguaje *Java*, el cual es uno de los más demandados al ser utilizado en diferentes plataformas de *software*. Se resaltaba la importancia de este curso exponiendo que los programadores *Java* son cada vez más buscados, por lo que es una característica básica que debería aparecer en el currículo de cualquier experto en tecnología.

El curso en cuestión fue la **primera parte de un bloque de tres MOOCs**. Esta primera parte apuntaba a familiarizar a los alumnos en los conceptos básicos, avanzando con conceptos más complejos a lo largo de las 5 semanas que duró. Aportaba ejemplos y casos de estudio para que los estudiantes pudieran implementar programas sencillos ellos mismos o colaborando con compañeros. La parte dos se centraría en cómo escribir “buenos” programas, y la tercera parte trataría sobre las estructuras de datos fundamentales (en estos momentos se puede referenciar la parte dos [32]).

Por último, se animaba de nuevo a apuntarse en el curso, mencionando que el conocimiento sobre programación no solamente es necesario para programar dispositivos actuales como ordenadores y *smartphones*, sino que también abre la puerta a la aplicación de técnicas computacionales a procesos rutinarios (*computational thinking*). En la versión del año 2016 del curso también aparecen algunas recomendaciones (*reviews*) escritas por alumnos del curso de 2015. La última vez que se comprobó, este *MOOC* poseía una calificación de 4,5 estrellas sobre 5 y tenía 74 *reviews*. Es interesante mencionar que, en la sección de recomendaciones, se puede observar si el usuario que ha escrito el análisis ha completado el curso, lo ha abandonado o lo está cursando en ese momento. También se muestra la puntuación individual que ha dado cada usuario.

En cuanto a los parámetros más técnicos del curso, éste tuvo una duración aproximada de 2 meses (aunque el contenido estaba concretamente dividido entre 5 semanas) y estimaba un esfuerzo de entre 5 y 7 horas por semana. Era un curso gratuito que además disponía de la posibilidad de recibir un certificado por 25 \$ (unos 22 €). Se incluyó en la categoría de Informática (*Computer Science*) a un nivel introductorio y se llevó a cabo en la Universidad Carlos III de Madrid (Departamentos de Ingeniería Telemática y de Informática de forma conjunta). El idioma en que se impartió fue inglés, y los vídeos utilizados tenían transcripciones en inglés, español, portugués y chino mandarín.

Volviendo a las 5 semanas en las que estaba organizado el curso, cada una de ellas contenía vídeos explicativos, textos, gráficas, tablas y ejercicios para que el alumno trabajase sobre ellos. En la versión que se estudia en este proyecto (la de 2015), existían *peer reviews* y los exámenes de evaluación tenían fechas fijas de finalización (dos

semanas después de ser liberados, excepto el primero, que tenía tres semanas). Efectivamente, esta política de fechas límite de exámenes disminuye el número de personas que aprueban el curso. Por el contrario, la versión de 2016 no poseía dichas actividades de revisión entre pares y los exámenes estuvieron disponibles hasta que acabó el curso. En cuanto al sistema de puntuación, hubo 5 **exámenes** de 15 puntos cada uno (uno por semana) y 2 actividades de *peer review* de 10 y 15 puntos en las semanas 3 y 5, respectivamente. Así, la puntuación podría variar entre 0 y 100, siendo necesaria una nota mínima de 60 puntos para aprobar el curso. Sobre las implicaciones de las tareas de revisión entre pares, se pueden mencionar algunas [33]. Por ejemplo, está el hecho de la escasez de supervisión por parte de los profesores debido a la gran cantidad de trabajos que evaluar (este problema sobre la alta cantidad de actividades también dificulta la coordinación para enlazar tareas y evaluadores). Otro hecho que se puede mencionar es la subjetividad de estas tareas y la escasa o nula experiencia en cuanto a la corrección de trabajos que puedan tener muchos participantes, lo cual puede desembocar en varias quejas sobre las puntuaciones obtenidas (algunos mensajes del foro y de *Twitter* tenían que ver con este tema). Cabe destacar que, en los cursos que tienen actividades de revisión entre pares, el índice de alumnos que completan dicho curso suele ser menor [33].

3.3 Foro y estructura de los mensajes

A continuación se hablará sobre los datos de los que se disponían al comenzar con este estudio y de cómo estaban éstos organizados. Primero, se tuvo acceso a los mensajes intercambiados en la cuenta de *Twitter* del curso. Éstos fueron recuperados utilizando *Ztreamey* [34], un prototipo de *software* destinado a publicar datos semánticos de la web (obtener datos legibles de aplicaciones informáticas). Más tarde, la propia organización *edX* facilitó un archivo de texto en el cual estaban presentes los **13778 mensajes** intercambiados en el foro de la asignatura, cada uno en una fila y con diferentes campos que se verán a continuación. De estos mensajes, 5897 se escribieron en discusiones ya creadas y el resto (7881) son inicios de nuevas discusiones (Figura 2). Dependiendo de si el mensaje pertenece al primer grupo o al segundo, tiene distintos campos, aunque algunos son comunes. En la siguiente tabla (Tabla 2) quedan apuntados dichos campos acompañados de una breve descripción del significado de cada uno [35]. Quedan resaltados los campos que se han utilizado en los algoritmos diseñados para la obtención de resultados.

Tipo	Campo	Descripción
Comunes a inicios de discusión y a comentarios en discusiones ya creadas	<code>_id</code>	Identificador único de cada mensaje, formado por números y letras
	<code>_type</code>	Es <i>CommentThread</i> en los inicios de discusión y <i>Comment</i> en los demás
	<code>anonymous</code>	Si se activa, la interfaz del usuario aparece como anónima para todos los demás miembros del curso, incluidos miembros del equipo de discusión (en este caso de estudio, 372 inicios de conversación tienen este campo marcado con <i>true</i>)

Comunes a inicios de discusión y a comentarios en discusiones ya creadas	<i>anonymous_to_peers</i>	Si se activa, la interfaz del usuario aparece como anónima solamente para los demás estudiantes
	<i>author_id</i>	Identificador numérico único del autor de cada mensaje
	<i>author_username</i>	Nombre de usuario del autor de cada comentario
	<i>body</i>	Contenido textual del mensaje
	<i>course_id</i>	En este caso de estudio, <i>UC3Mx/IT.1.1x/IT2015</i>
	<i>created_at</i>	Marca de tiempo de la creación del mensaje en formato UTC
	<i>updated_at</i>	Marca de tiempo de la actualización del mensaje en formato UTC
	<i>votes</i>	Número de estudiantes que han valorado positiva y/o negativamente un comentario, al igual que sus respectivos IDs
Inicios de discusión	<i>closed</i>	Indica si el hilo fue cerrado por un moderador o administrador (en este caso de estudio, todos los inicios tienen este campo seguido de <i>false</i>)
	<i>comment_count</i>	Número de comentarios respuesta al inicio de discusión
	<i>commentable_id</i>	Indica si la discusión está adjunta a algún contenido del curso
	<i>last_activity_at</i>	Marca de tiempo del último instante en el que hubo actividad en el hilo en formato UTC
	<i>title</i>	Título textual de la discusión
	<i>thread_type</i>	Puede ser <i>discussion</i> o <i>question</i> (en este caso de estudio, existen 7112 hilos del primer tipo y 769 del segundo; 7881 en total)
Comentarios en discusiones ya iniciadas	<i>visible</i>	Campo presente pero no usado
	<i>abuse_flaggers</i>	IDs de los usuarios que denuncian (<i>Report Misuse</i>) este comentario
	<i>historical_abuse_flaggers</i>	Se mueve a este campo el contenido de <i>abuse_flaggers</i> cuando un moderador elimina la marca <i>Report Misuse</i>
	<i>endorsed</i>	Si es <i>true</i> , un moderador (típicamente, el profesor) ha marcado este comentario como una valiosa contribución en un hilo de tipo <i>discussion</i> o como una respuesta correcta en uno de tipo <i>question</i> (en este caso de estudio, 65 mensajes poseen campo seguido de <i>true</i>)
	<i>endorsement</i>	Marca de tiempo e ID del usuario que marcó el comentario como <i>endorsed</i>

Comentarios en discusiones ya iniciadas	<i>comment_thread_id</i>	Identifica a qué inicio de discusión corresponde este comentario
	<i>parent_id</i>	Si se está respondiendo a otro comentario, es el ID de éste (nunca el ID de un inicio de discusión)
	<i>parent_ids</i>	IDs de todos los antecesores del mensaje (debido a la interfaz de usuario actual, solamente muestra como mucho un ID)
	<i>sk</i>	Número aleatorio para mejorar la indexación <i>online</i>
Tipo	Campo	Descripción

Tabla 2. Estructura de los mensajes del foro.

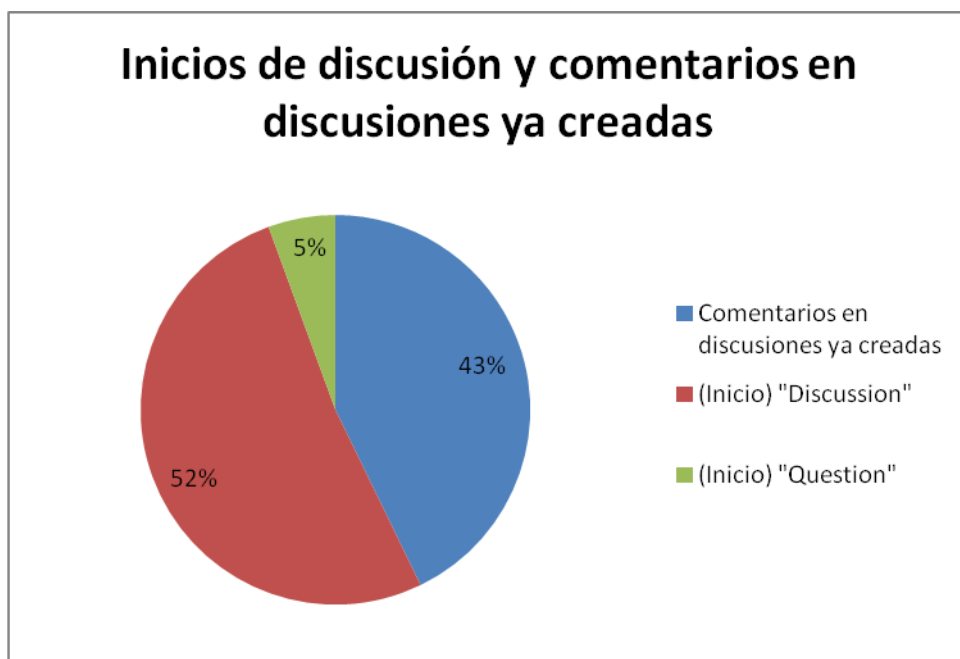


Figura 2. Inicios de discusión y comentarios en discusiones ya creadas.

3.4 Twitter

Como se mencionó anteriormente, al comienzo de este proyecto se tuvo acceso a un archivo que poseía los *tweets* que contenían el *hashtag* del curso, cada uno con diversos campos. Se realizó un recuento de los mismos atribuyéndolos a sus respectivos autores, aunque su estudio prácticamente no ha tenido relevancia en este trabajo. Como **características a resaltar**, en total se escribieron 197 *tweets* entre 98 usuarios diferentes, incluyendo coordinadores y *staff*. Tal vez cabría destacar que la cuenta que más mensajes compartió (36) fue la del propio curso (*javaedxuc3m*), seguida por otro usuario (*Tundeiness*, con 13 *tweets*) que ocupa el puesto 53 en cuanto a los mayores contribuidores al foro. Aún así, el tercer usuario que escribió más *tweets* (*hugomesmo*, con 7 mensajes) sí está dentro del *top 10* de contribuidores al foro del curso. En la Tabla 3 se ha realizado una clasificación general de todos los mensajes.

Tipo	Cantidad	Porcentaje
Difusión/Saludos	94	47,7
Positivos	16	8,1
Negativos	6	3,1
Neutros	15	7,6
Preguntas/Problemas	40	20,3
Respuestas	26	13,2

Tabla 3. Clasificación de los *tweets*.

Lo más destacable es el alto porcentaje de mensajes destinados a la difusión del *MOOC*. La mayoría de estos *tweets* eran anuncios escritos por la propia cuenta del curso, otros eran *retweets* de éstos y algunos otros usuarios decían que se habían inscrito en el *MOOC*. En cuanto a las preguntas, unas eran sobre la obtención del certificado y otras estaban relacionadas con las actividades de evaluación, mayormente.

En realidad, no es sorprendente que esta red social sirva más como **comunicación o difusión del curso** que como mecanismo para que los usuarios interactúen entre ellos. Por ejemplo, la restricción de que un *tweet* solamente pueda contener 140 caracteres dificulta considerablemente conversaciones de tipo “preguntas y respuestas”. De hecho, la cuenta del curso respondió a un usuario que había formulado una pregunta sugiriéndole que escribiera una entrada más detallada en el foro de discusión.

Capítulo 4

Diseño e implementación de la solución

4.1 Diseño de la solución

Como ya se dijo en el Capítulo 3, el objetivo general que se plantea es la obtención de información relevante a partir de los datos que se pudieran sacar del archivo que contiene las contribuciones al foro. Dicha información servirá tanto para comparar los resultados con las hipótesis y conclusiones planteadas en otros documentos como para investigar las características de los *top contributors*. Para ello, y como también se sugiere brevemente en el capítulo citado, fue necesario **automatizar** de alguna forma la recogida de los datos y la generación de resultados. Con todo, podrían diferenciarse varias fases en este proceso.

Lo primero que se debía realizar consistía en identificar una nueva línea de investigación; esto es, qué información se desea obtener. Segundo, de acuerdo con la fase anterior, era necesario un proceso de organización de los datos de partida en un archivo *Excel*. A continuación venía la fase de lectura y procesado de estos datos a través de la herramienta *MATLAB* y mediante la programación de algoritmos con este *software*. Finalmente, dichos algoritmos proporcionarían un resultado y la fase final consistiría en estructurar e interpretar estos resultados. Efectivamente, el proceso compuesto por estas fases se realizaba de **forma iterativa** cada vez que surgía una idea nueva a partir de los resultados previos que se iban obteniendo en el proyecto.

4.2 Programas diseñados

Como se ha mencionado, se utilizó el programa *MATLAB* para la realización y ejecución de los algoritmos programados. Estos archivos tienen la extensión *.m* y el lenguaje que utiliza la aplicación se puede considerar parecido al lenguaje *C* o *Java*. A continuación se muestra una tabla (Tabla 4) que identifica el nombre de los *scripts* (“secuencias de comandos”) utilizados junto con una breve descripción de la misión de cada uno.

Nombre	Tiempo medio de ejecución	Líneas de código	Descripción
<i>Certificados</i> (Algoritmo 1)	20 s	123	Calcula el número y porcentaje de alumnos aprobados y suspensos en dos grupos según comentaron o no en el foro (Figura 13)
<i>Certificados_2semanas</i> (Algoritmo 1)	15 s	116	Similar a <i>Certificados</i> pero solamente para los datos pertenecientes a las dos primeras semanas del curso (Figura 14)
<i>ClasificacionDiagramaSocial</i> (Algoritmo 2)	174 s (2,9 min)	49	Clasifica todos los mensajes en pares “respondedor-respondido”
<i>Histogramas</i> (Algoritmo 3)	834 s (13,9 min)	115	Tanto para todo el curso como para las dos primeras semanas, representa un histograma de “número de mensajes” frente a “usuarios con X número de mensajes” (Figura 8) También se divide el histograma en intervalos de distinta escala (Figura 9 y Figura 10)
<i>RecuentoCodigo</i> (Algoritmo 4)	Op. 1: 576 s (9,6 min) Op. 2: 576 s (9,6 min) Op. 3: 14 s Op. 4: 6 s	176	Permite elegir entre 4 opciones Opción 1: cuenta el número de comentarios publicados por cada usuario, identificando a éste con su <i>author_username</i> Opción 2: cuenta el número de comentarios publicados por cada usuario, identificando a éste con su <i>author_id</i> Opción 3: cuenta el número de mensajes que fueron publicados cada día (Figura 7) Opción 4: cuenta el número de mensajes que fueron publicados en cada franja horaria para todo el curso y eliminando los tres primeros días (Figura 11 y Figura 12, respectivamente)

<i>RecuentoCodigo2semanas</i> (Algoritmo 4)	Op. 1: 255 s (4,25 min) Op. 2: 258 s (4,3 min)	27	Similar a <i>RecuentoCodigo</i> pero restringido a las dos primeras semanas del curso y sin la Opción 3 ni la Opción 4
<i>RestriccionDiagramaSocial</i> (Algoritmo 5)	4 s	25	A partir de los resultados de <i>ClasificacionDiagramaSocial</i> y del número de comentarios de cada usuario, realiza otra clasificación “respondedor-respondido” restringiéndola a usuarios que hayan aportado más de un número especificado de contribuciones al foro (Figura 23 y Figura 24)
<i>UltimoComentarioAprobados</i> (Algoritmo 6)	17 s	138	De los alumnos que aprobaron el curso, selecciona la fecha de su última publicación en el foro (se puede variar un parámetro para hacer lo mismo con los usuarios suspensos; Figura 15, Figura 16 y Figura 17)
<i>PlotsTops</i> (Algoritmo 7)	7 s	71	De los 20 usuarios más activos, cuenta el número de aportaciones que realizó cada uno en cada día del curso (Figura 18, Figura 19, Figura 20 y Figura 21)
<i>Correlaciones</i> (Algoritmo 8)	3,9 s	48	Calcula la correlación entre la nota obtenida por los alumnos (eliminando a los que sacaron un 0) y el número de comentarios que éstos publicaron También calcula la correlación entre la nota y una relación establecida entre el número de discusiones iniciadas y el número de comentarios en discusiones ya creadas
<i>Notas</i> (Algoritmo 9)	12 s	32	Para todo el rango de notas posibles (eliminando el 0), calcula el número de estudiantes que obtuvieron cada una (Figura 6)
Nombre	Tiempo medio de ejecución	Líneas de código	Descripción

Tabla 4. *Scripts* de *MATLAB* creados

Tras esta tabla, a continuación se mostrarán los pasos que sigue cada *script* mediante la utilización de pseudocódigo.

CAPÍTULO 4: DISEÑO E IMPLEMENTACIÓN DE LA SOLUCIÓN

Algoritmo 1. *Certificados* (también aplicable a *CertificadosDosSemanas*; lo único que cambian son los datos leídos al comienzo del código)

```
READ document about certificates
READ number of contributions of each user
TRANSFORM string data into numeric data
FOR each user that contributed to the forum
    IF user is found in the document about certificates THEN
        Assign his/her final mark on the course
    ELSE
        INCREMENT the number of missing users
    ENDIF
ENDFOR
DETERMINE order by number of contributions
DETERMINE data eliminating teachers and staff
CALCULATE number and percentage of students passing and failing
the course
CALCULATE number and percentage of contributors passing and
failing the course
CALCULATE number and percentage of no-contributors passing and
failing the course
COMPUTE data and graphs about users passing the course with more
than a certain number of contributions
```

Algoritmo 2. *ClasificacionDiagramaSocial*

```
READ user IDs of creators of comment threads
READ IDs of comment threads
READ user IDs of contributors to already created threads
READ IDs of parent comments
FOR each user that commented on already created threads
    DETERMINE user ID
    DETERMINE which threads the user contributed to
    FOR each thread the user contributed to
        STORE ID of the user who replied and ID of the thread
creator
    IF the comment has another parent comment THEN
        STORE ID of the user who replied and ID of the parent
    ENDIF
    IF another ten percent of the data has been processed THEN
        DISPLAY progress
    ENDIF
ENDFOR
```

Algoritmo 3. *Histogramas*

```
READ username of the creator of each contribution
FOR each unique username
    COMPUTE the number of contributions of this user
    IF another ten percent of the data has been processed THEN
        DISPLAY progress
    ENDIF
ENDFOR
DISPLAY graph
```

```

READ username of the creator of each contribution in the first
two weeks of the course
FOR each unique username
    COMPUTE the number of contributions of this user
    IF another ten percent of the data has been processed THEN
        DISPLAY progress
    ENDIF
ENDFOR
DISPLAY graph
DISPLAY graph of the whole course dividing data in subplots with
specified intervals
DISPLAY graph of the first two weeks dividing data in subplots
with specified intervals

```

Algoritmo 4. *RecuentoCodigo* (también aplicable a *RecuentoCodigo2semanas* eliminando la opción tres)

```

DISPLAY list of options
IF option 1 is selected THEN
    READ username of the creator each comment
    FOR each unique username
        COMPUTE the number of contributions of this user
        STORE data
        IF another ten percent of the data has been processed
THEN
            DISPLAY progress
        ENDIF
    ENDFOR
ELSEIF option 2 is selected THEN
    READ user ID of the creator each comment
    FOR each unique user ID
        COMPUTE the number of contributions of this user
        STORE data
        IF another ten percent of the data has been processed
THEN
            DISPLAY progress
        ENDIF
    ENDFOR
ELSEIF option 3 is selected THEN
    GENERATE date "2015-04-28"
    WHILE not all the messages have been processed
        COMPUTE the number of comments posted on the current
date
        STORE data
        INCREMENT date by 1 day
        IF another ten percent of the data has been processed
THEN
            DISPLAY progress
        ENDIF
    ENDWHILE
ELSEIF option 4 is selected THEN
    GENERATE hour "0"
    WHILE not all hours have been processed
        COMPUTE the number of comments posted on the current
hour for the entire course

```

CAPÍTULO 4: DISEÑO E IMPLEMENTACIÓN DE LA SOLUCIÓN

```
        COMPUTE the number of comments posted on the current
hour without taking into account the first three days

        STORE data
        INCREMENT hour by 1
        IF another five percent of the data has been processed
THEN
        DISPLAY progress
    ENDIF
ENDWHILE
DISPLAY graphs
ELSE
    DISPLAY list of options
ENDIF
```

Algoritmo 5. *RestriccionDiagramaSocial*

```
READ user IDs and the number of contributions of each one
READ user IDs of the ones that replied and user IDs of the ones
being replied
SELECT number of minimum comments
FOR each user that contributed with at least the minimum number
of comments
    DETERMINE the user's position within the data containing
users that replied
ENDFOR
DELETE users that did not contribute with at least the minimum
number of comments
STORE data
```

Algoritmo 6. *UltimoComentarioAprobados*

```
READ user ID of the contributors that passed the course
READ the dates of creation of their comments
GENERATE date "2015-04-28"
WHILE the month of the date is lesser than 12
    DETERMINE which dates of creation match this date and
rewrite them with another format
    INCREMENT date by 1 day
ENDWHILE
READ user IDs, number of comments and final mark of all users
FOR each user that passed/failed the course
    STORE the user ID
ENDFOR
FOR each unique user ID that passed/failed
    DETERMINE the dates in which the user commented
    STORE the last one
ENDFOR
MODIFY data types to assign user ID with date of last comment
```

Algoritmo 7. *PlotsTops*

```
READ user ID and date of each comment
REWRITE dates with a new format
SELECT user IDs of the top contributors
```

```

FOR each top contributor's user ID
    DETERMINE the dates in which the user commented
    FOR each date in which the user commented
        COMPUTE the number of comments in each date
        STORE data
    ENDFOR
ENDFOR
FOR each top contributor's user ID
    SELECT the dates in which the user commented and the number
of comments each date
    FOR each day of the duration of the course
        FOR each date in which the user commented
            IF both day of the course and date of comment
coincide THEN
                STORE date and number of comments
            ENDIF
        ENDFOR
    ENDFOR
ENDFOR

```

Algoritmo 8. Correlaciones

```

READ mark, number of comments and ID of each user
DELETE users whose mark is 0
COMPUTE the correlation between mark and number of comments
DISPLAY result
READ user IDs of participant that started discussions and user
IDs of participants that commented in already created discussions
FOR each user ID
    COMPUTE number of started discussions and number of comments
in already created discussions
    IF both number of started discussions and number of comments
in already created discussions are 0 THEN
        CONTINUE with next user
    ENDIF
    STORE ID, number of started discussions, number of comments
in already created discussions and mark
ENDFOR
DELETE users whose mark is 0
COMPUTE specified relation between number of started discussions
and number of comments in already created discussions
COMPUTE correlation
DISPLAY result

```

Según los resultados devueltos por este algoritmo, la correlación entre la nota obtenida y el número de comentarios fue de 0,20761. Por otro lado, tras probar diferentes combinaciones entre número de inicios de discusión y cantidad de aportaciones a hilos ya creados, se utilizó la siguiente expresión para calcular una segunda correlación:

$$\frac{C}{C+I} \begin{cases} C = \text{número de comentarios en discusiones ya creadas} \\ I = \text{número de inicios de discusión} \end{cases}$$

Así, la correlación obtenida entre la nota final de los usuarios y la anterior expresión fue de 0,2538. Este valor resultó ser el más alto de entre todas las relaciones probadas.

Algoritmo 9. *Notas*

```
READ mark of each student
REWRITE marks with a new format
DELETE zeros
FOR each unique mark
    COMPUTE the number of times that mark appears
ENDFOR
DISPLAY graph
```


Capítulo 5

Resultados²

5.1 Curso

Primero se apuntarán los resultados que afectan de manera general al curso. A continuación se mostrarán los **resultados numéricos de las acciones de los usuarios en torno al foro de discusión**. Se tiene que, eliminando a los profesores y al *staff*, aprobaron 1522 alumnos y suspendieron 83246 de un total de 84768 alumnos (Figura 3). Entre los que suspendieron, 76320 obtuvieron como nota final un cero (90% sobre todos los alumnos; 91,7% sobre los alumnos que suspendieron). Estos valores pueden resultar relevantes ya que sesgan mucho los resultados. Adicionalmente, se aportan más datos dependiendo de si el estudiante hizo aportaciones al foro o no. Hubo 79538 alumnos que no comentaron en el foro. De ellos, 764 aprobaron el curso (0,96% de los que no participaron). Por otra parte, de los 5230 que sí comentaron, 758 aprobaron el curso (14,49% de los que sí participaron). Además, se puede realizar un agrupamiento más en función de si el usuario aprobó o suspendió. De todos los que aprobaron (1522), 764 alumnos no realizaron aportaciones (50,2%) y los 758 restantes sí las realizaron (49,8%; Figura 4). Por otro lado, de los que suspendieron (83246), 78774 no realizaron comentarios (94,63%) y los 4472 alumnos restantes sí comentaron (5,37%; Figura 5). De estos resultados tal vez quepa destacar brevemente que la mitad de los alumnos que aprobaron no realizaron aportaciones al foro. Éste es un resultado muy curioso a la hora de comprobar la correlación entre aprobar el curso y aportar en el foro. También es notable que el 98% de todos los alumnos suspendieron y que el 90% del total lo hizo con

² Los resultados específicos de la red social *Twitter* ya fueron apuntados al final del Capítulo 3.

un cero por nota final, lo cual da una idea de la cantidad de alumnos que no se involucran en el curso (por ejemplo, participantes que ven los vídeos y los documentos pero no realizan las actividades) o que sencillamente se apuntan y no lo siguen.

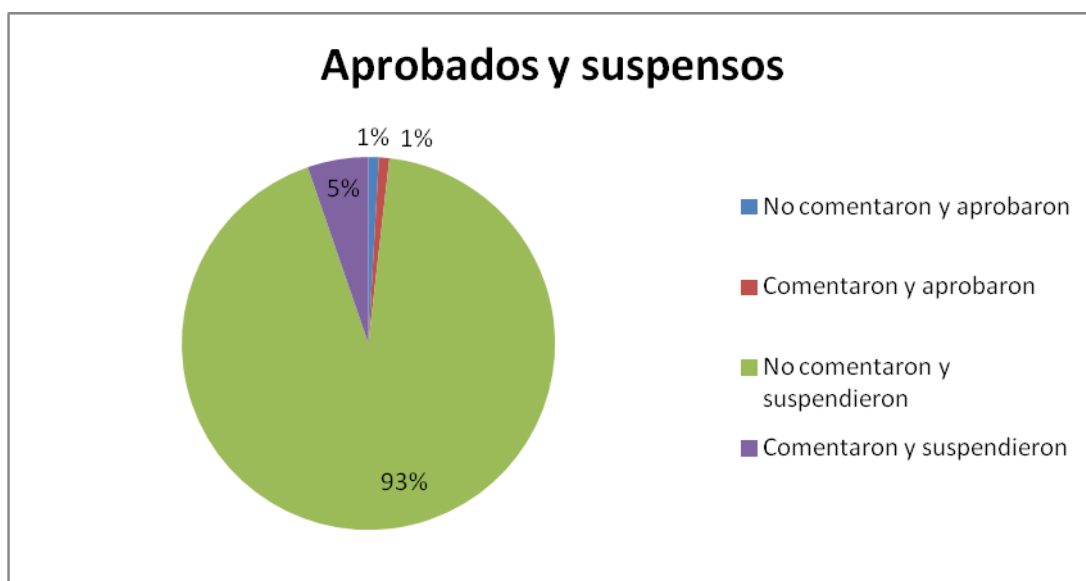


Figura 3. Porcentaje de aprobados y suspensos.

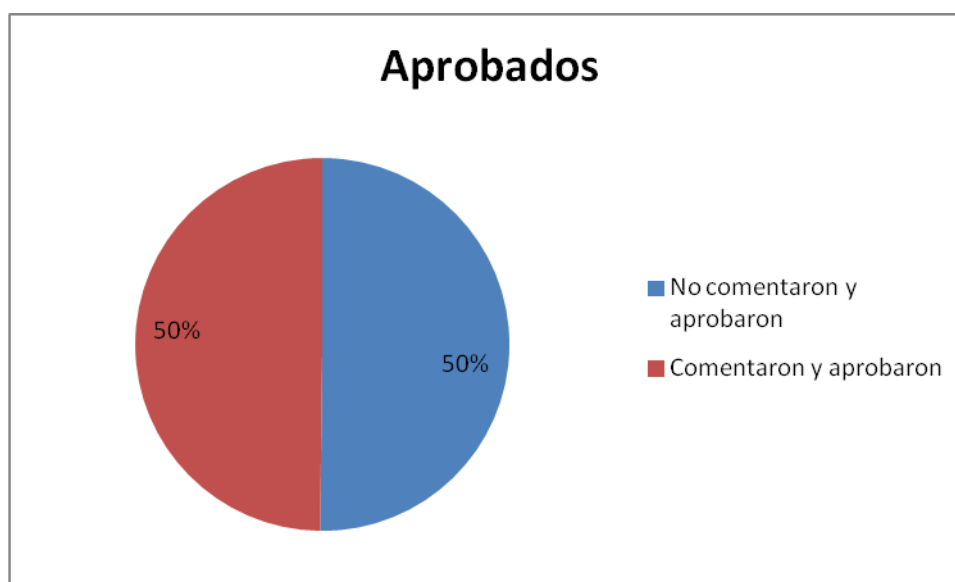


Figura 4. Porcentaje de aprobados que comentaron y no comentaron en el foro de discusión.

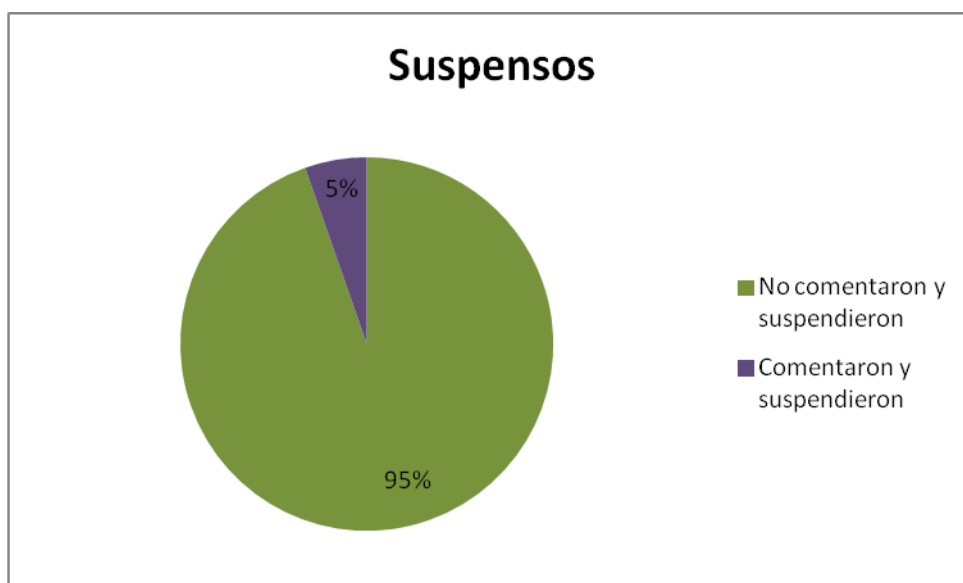


Figura 5. Porcentaje de suspensos que comentaron y no comentaron en el foro de discusión.

En relación con esta clasificación de aprobados y suspensos se puede contemplar la **nota que sacaron los participantes** (Figura 6). Se han excluido los 76320 alumnos (90% del total) que tuvieron un cero como nota final. Es relevante la alta cantidad de alumnos con una puntuación menor que 16. Como se comentó anteriormente, con cada examen se podía obtener un máximo de 15 puntos. Así, resulta interesante que de esta gráfica se pueda deducir que muchos alumnos solamente hacen un examen (típicamente, el primero).

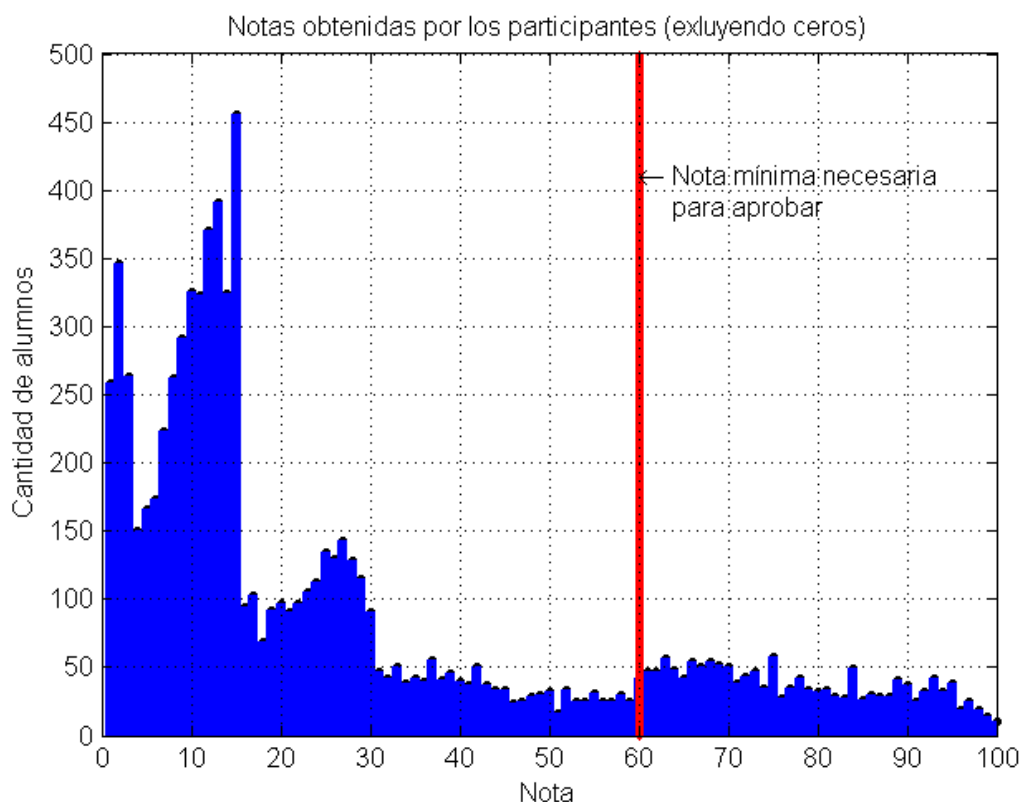


Figura 6. Notas obtenidas por los participantes excluyendo a los que sacaron un cero. Nota: no se han excluido a los coordinadores y *staff*.

Una vez vistos estos datos generales sobre los participantes, los resultados que se mostrarán a partir de ahora solamente conciernen a los **alumnos que sí realizaron alguna aportación** al foro de discusión. En la Figura 7 se muestra la evolución del número de comentarios según avanzaba el curso (el rango de fechas varía desde el inicio del *MOOC* hasta la fecha del último comentario registrado). Es destacable el pico inicial provocado por los mensajes de bienvenida y presentaciones; después se tiene una media de unos 200 mensajes al día hasta la finalización del curso. Los demás picos suelen coincidir con liberaciones de material, lo cual genera mucha discusión en esos mismos días (por ejemplo, los 306 mensajes del día 2 de junio, última fecha en la que se liberan materiales nuevos). La gráfica obtenida a partir de los datos sigue una distribución similar a la obtenida en el estudio del curso “Educación Digital del Futuro” [13].

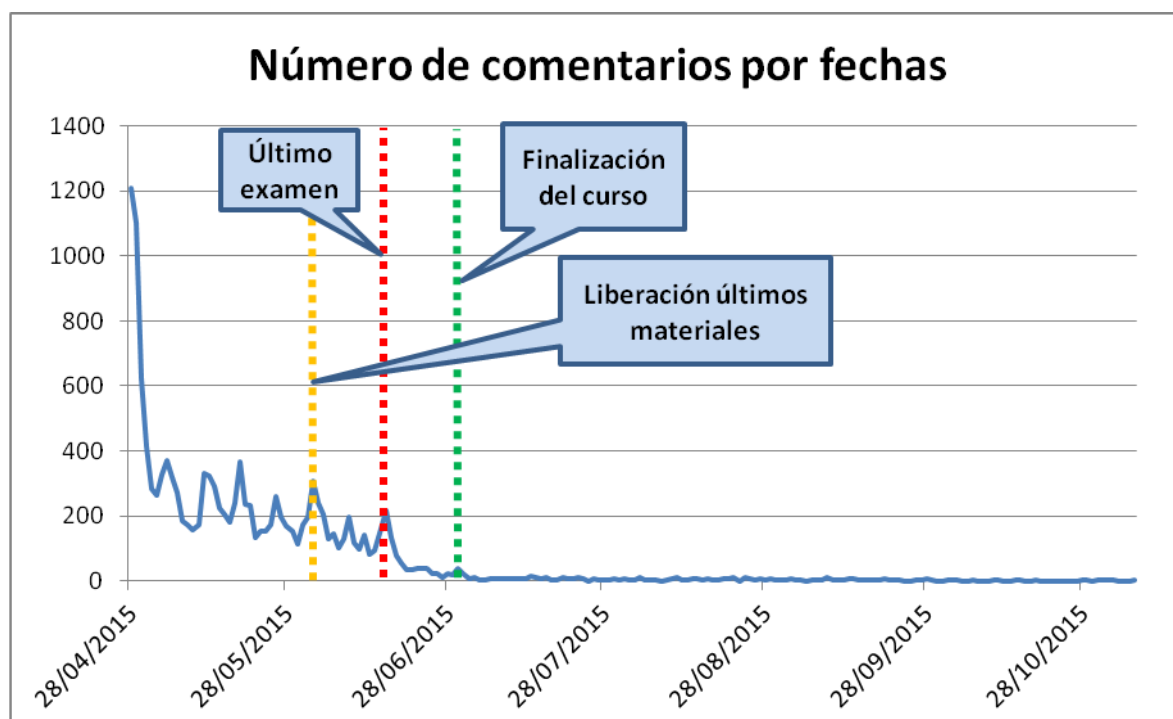


Figura 7. Número de comentarios por fechas.

Siguiendo con recuentos de contribuciones, se ha calculado el **número de alumnos que escribieron distintas cantidades de mensajes** en el foro (Figura 8). Como se puede ver, la gran mayoría solamente realizó una aportación, la cual en muchos casos coincide con el periodo de bienvenida del curso. La media de todo el curso es de 2,5 mensajes por usuario (de entre los que realizaron alguna aportación).

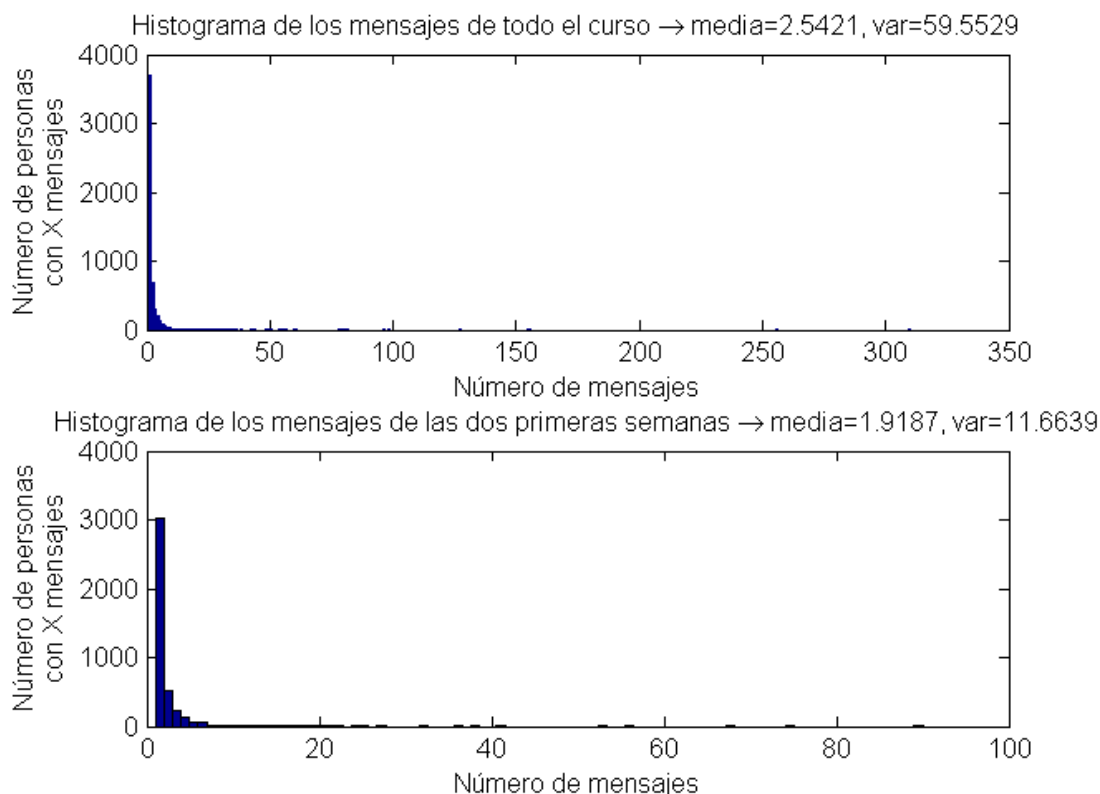


Figura 8. Histogramas representando el número de usuarios que realizaron diferentes cantidades de aportaciones al foro de discusión.

Algunos de los cálculos, como los histogramas anteriores, se han realizado también para las **dos primeras semanas del curso**. Este periodo es importante por algunos motivos. Por una parte, en los primeros días se concentra la emoción inicial por el comienzo del curso, como ya se mencionó anteriormente. Por otra parte, hasta que no pasaron dos semanas, no se liberaron los materiales de la “segunda semana” oficial del curso. También se podría citar la importancia de este periodo para intentar identificar de manera temprana a los mayores contribuidores. Durante estas dos primeras semanas, se tiene que 80592 alumnos no comentaron en el foro. De ellos, 1011 aprobaron el curso (1,25% de los que no contribuyeron). Por otro lado, de los 4176 que sí comentaron, 511 aprobaron el *MOOC* (12,24% de los que sí contribuyeron). Para una mejor visualización de los histogramas (sobre todo para apreciar los valores más bajos), se han dividido éstos entre intervalos con diferentes longitudes (Figura 9 y Figura 10).

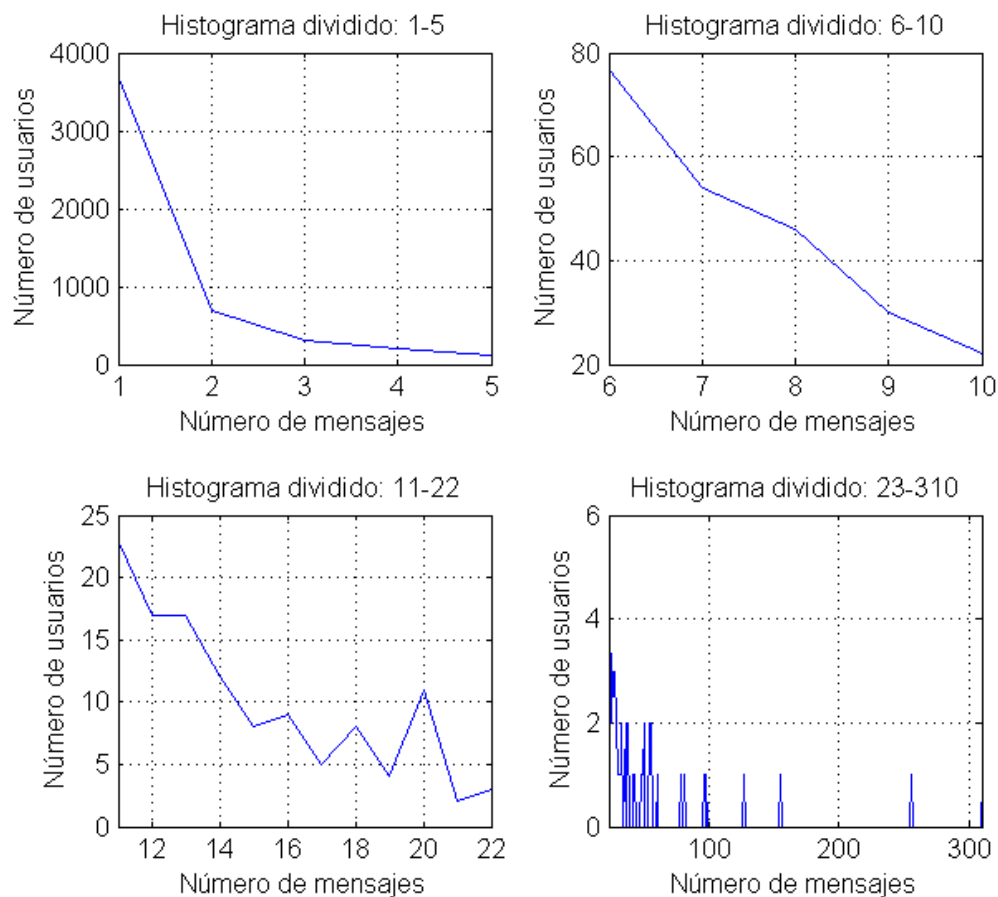


Figura 9. Histograma dividido en intervalos con diferentes longitudes para una mejor visualización (todo el curso). Nota: no se han excluido a los coordinadores y *staff*.

Efectivamente, estos histogramas sufren su mayor caída entre uno y dos comentarios. A partir de ese momento, las gráficas en general van decreciendo, hasta el momento de encontrarse con puntos aislados a partir de los 25 comentarios aproximadamente.

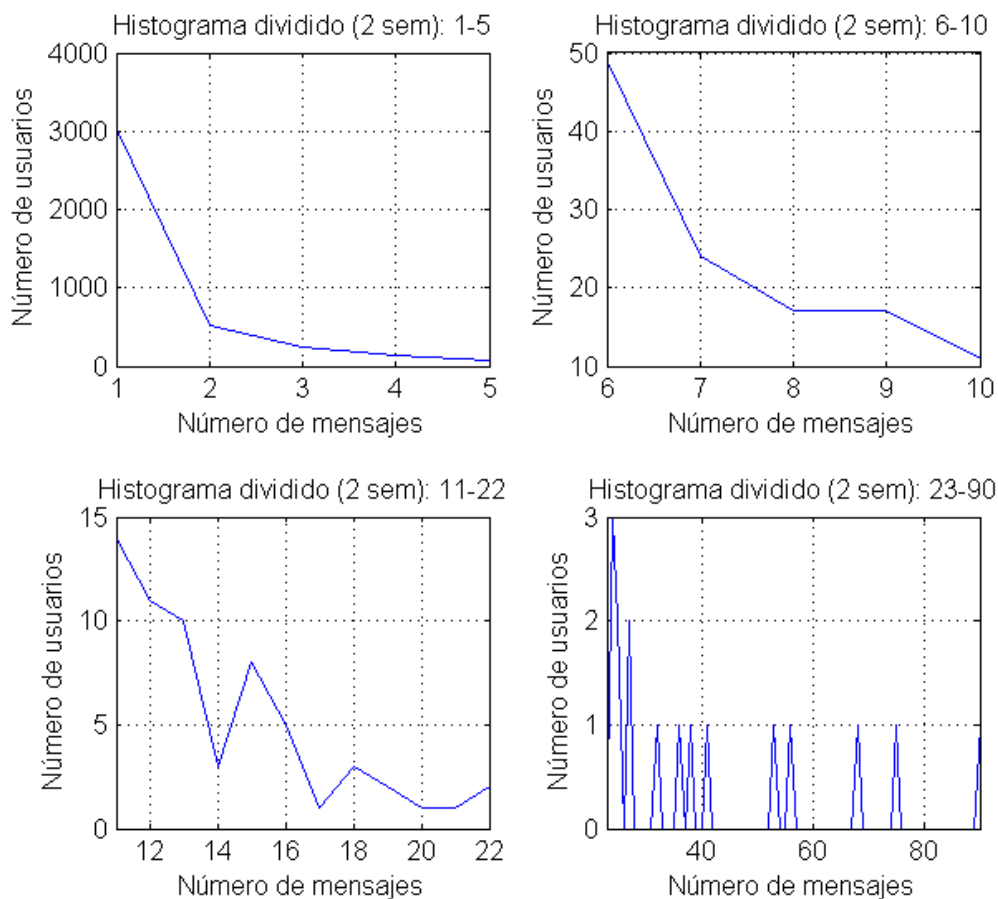


Figura 10. Histograma dividido en intervalos con diferentes longitudes para una mejor visualización (primeras dos semanas de curso). Nota: no se han excluido a los coordinadores y staff.

Se ha visto una clasificación de la cantidad de mensajes registrados según fechas; también se realizó una clasificación similar en función de la **franja horaria** en la que fueron realizadas las publicaciones (Figura 11 y Figura 12). En el fichero utilizado, las horas aparecían en formato UTC [36]. Las gráficas reflejan la tarde temprana europea (UTC+02:00) y también la tarde de Estados Unidos (entre UTC-09:00 y UTC-05:00), pero la tarde de India (UTC+5:30) tiene menor impacto.

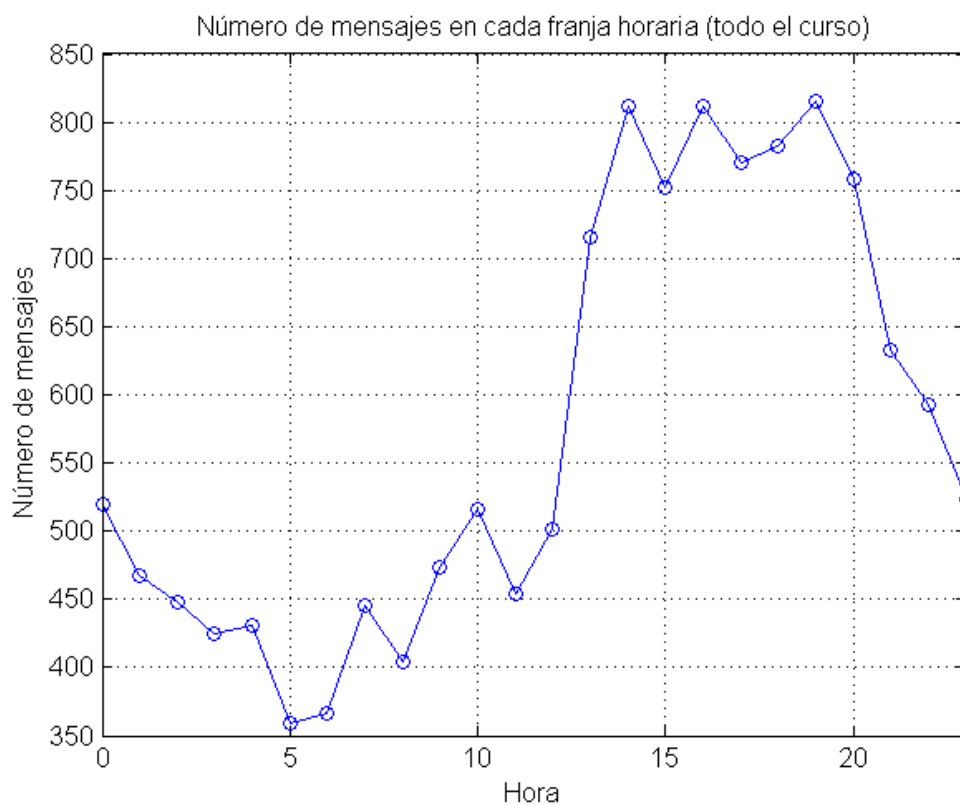


Figura 11. Número de mensajes en cada franja horaria (todo el curso; formato UTC). Nota: el valor que corresponde a una hora, X, es el número de mensajes publicados entre X:00:00 y X:59:59.

En la Figura 12 se ha representado lo mismo dejando fuera de los cálculos los **tres primeros días** del curso para intentar no sesgar los resultados con los mensajes iniciales de bienvenida; estos tres primeros días contienen el 21,3% de los mensajes intercambiados durante todo el curso. Ambas gráficas siguen una distribución similar.



Figura 12. Número de mensajes en cada franja horaria (eliminando los 3 primeros días; formato UTC). Nota: el valor que corresponde a una hora, X, es el número de mensajes publicados entre X:00:00 y X:59:59.

Retomando de nuevo la clasificación entre los alumnos que aprobaron y los que no, anteriormente se vio el número y porcentaje de participantes que habían aprobado comentando en el foro alguna vez y los que aprobaron sin comentar nunca. Yendo un paso más lejos, se ha representado el número y porcentaje de estudiantes que, habiendo realizado al menos **cierta cantidad de aportaciones al foro, aprobaron** el curso (Figura 13). Se puede resaltar que aproximadamente desde el 30% de aprobados hasta el 100% la gráfica aumenta de forma relativamente lineal con el número de comentarios realizados.

CAPÍTULO 5: RESULTADOS

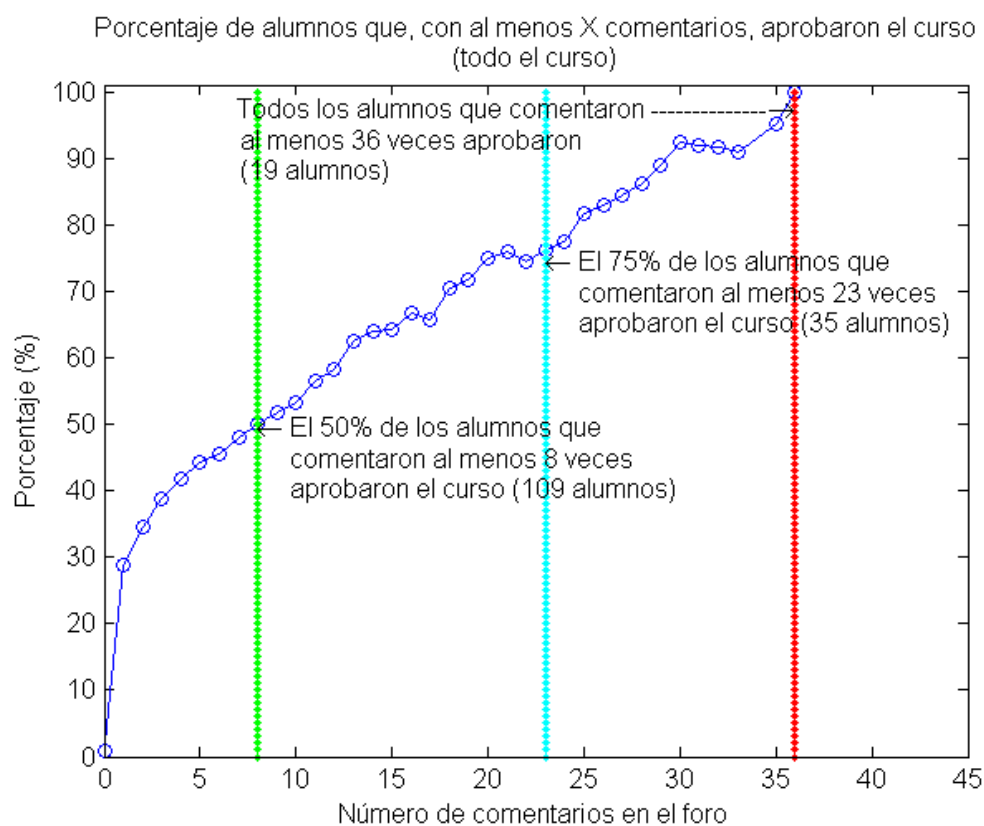


Figura 13. Para todo el curso, porcentaje de alumnos que, con al menos un número dado de aportaciones al foro de discusión, aprobaron el MOOC.

Igualmente, se realizaron los mismos cálculos restringiendo éstos a las dos primeras semanas del curso (Figura 14). En este caso, la aproximación a una recta es menos fiable. Cabría esperar que los alumnos que se han involucrado en el curso durante las dos primeras semanas seguirán adelante con éste: por ejemplo, la mitad de los alumnos que realizaron más de 17 aportaciones al foro durante este periodo aprobaron el MOOC.

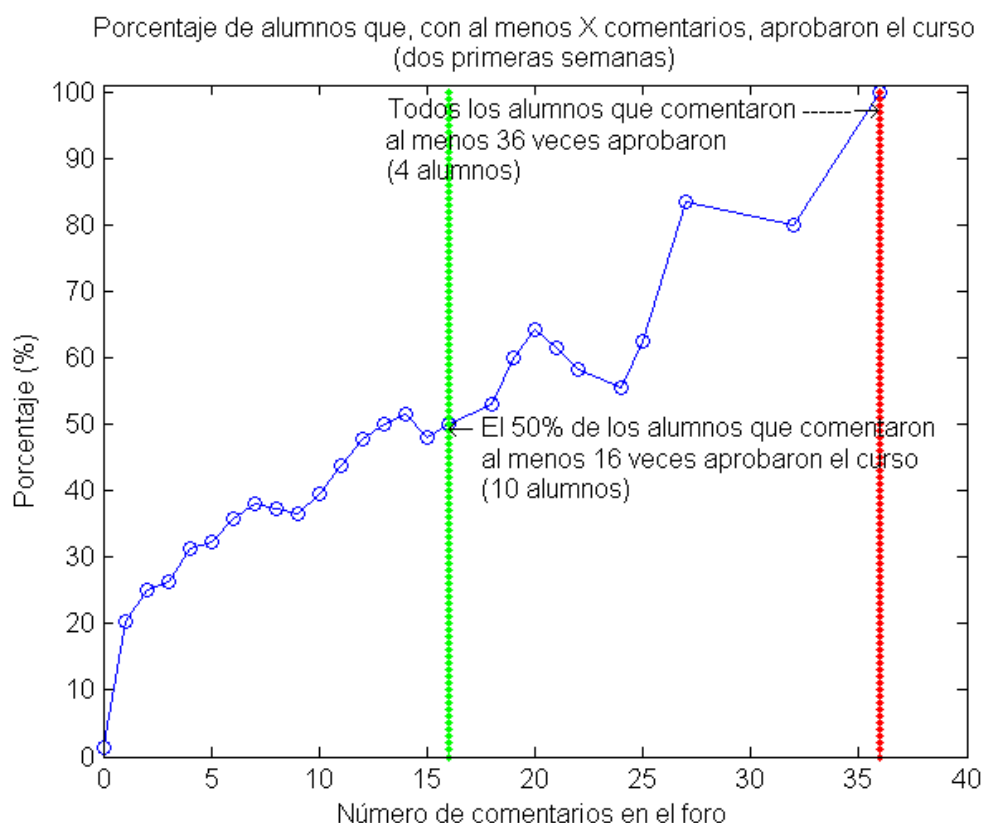


Figura 14. Para las dos primeras semanas del curso, porcentaje de alumnos que, con al menos un número dado de aportaciones al foro de discusión, aprobaron el MOOC.

Los datos obtenidos hasta ahora aportan una idea de a partir de cuántas aportaciones se puede inferir, con mayor o menor seguridad, que un alumno haya aprobado el curso (o que vaya a aprobar, si se tiene en cuenta la Figura 14). Ahora bien, se debe notar que solamente introducían una referencia temporal los datos que contemplaban los números de comentarios por fechas (Figura 7). Los siguientes datos también reflejarán esta **referencia temporal**. Primero, se puede observar, respecto a los alumnos que suspendieron el curso, cuándo fue la fecha de su último comentario (Figura 15). No es sorprendente el pico inicial de alumnos que sufre la gráfica; se puede deducir que la mayoría de esos alumnos comentaron y saludaron por la emoción inicial del curso pero no siguieron con él.

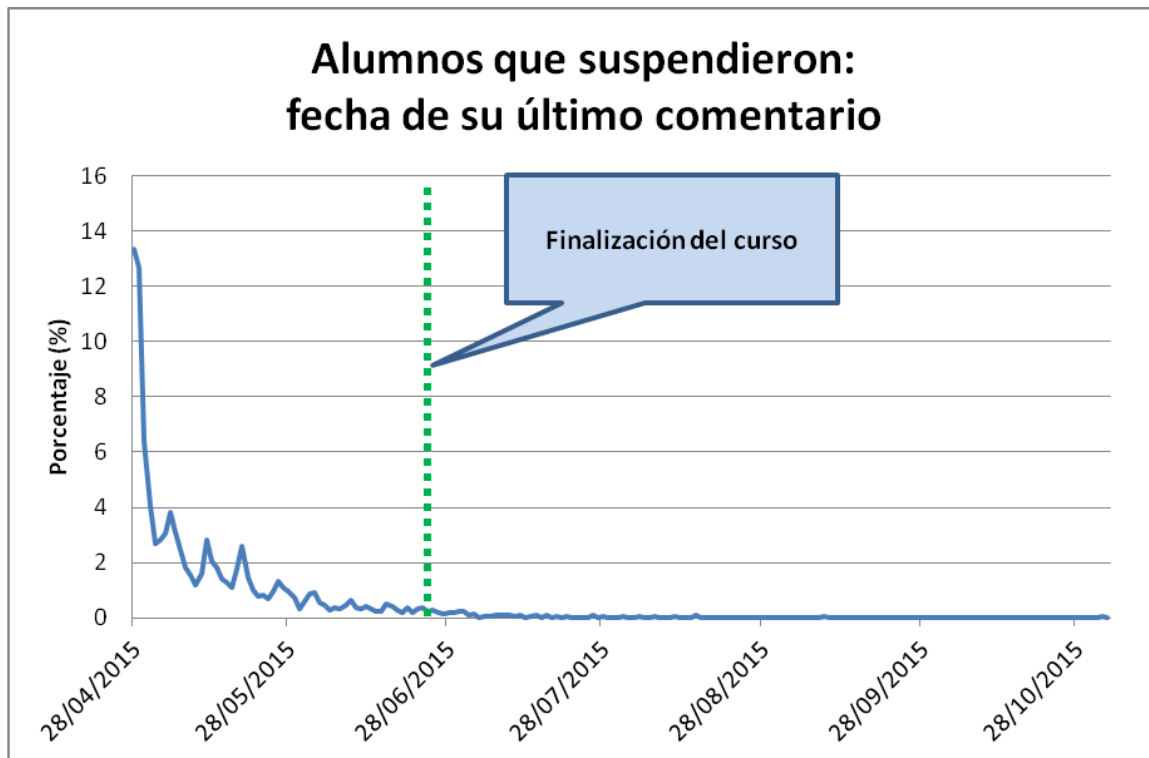


Figura 15. Fechas de publicación de la última aportación al foro de los alumnos que suspendieron.

Asimismo, se puede obtener otra gráfica sustituyendo a los alumnos que suspendieron por los que aprobaron (Figura 16). Destaca el hecho de que en este caso también haya un pico inicial en torno a los saludos y las discusiones de bienvenida. Ello significa que un porcentaje considerable de los alumnos que aprobaron y comentaron solamente realizaron aportaciones los primeros días. Aún así, a diferencia del caso de los alumnos suspensos, éstos siguieron con el curso y tal vez consultaran el foro aunque no escribieran en él. También es notable que un gran número de alumnos siga comentando hasta más de la mitad del *MOOC*, y que dejen de hacerlo en fechas cercanas a la finalización del mismo.

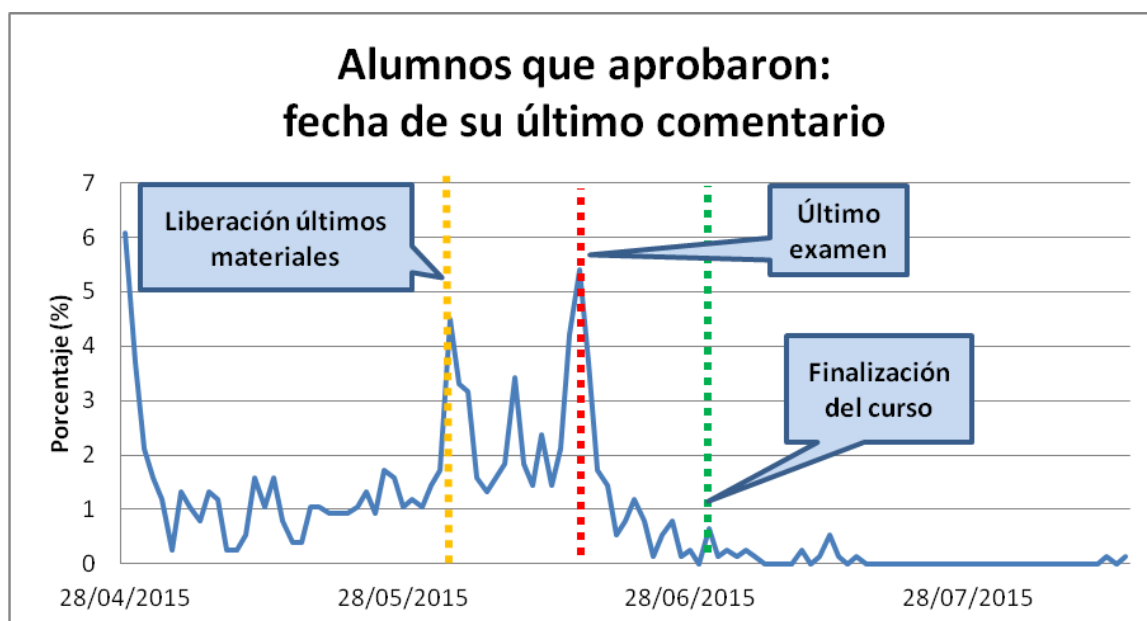


Figura 16. Fechas de publicación de la última aportación al foro de los alumnos que aprobaron.

Otra forma de observar los datos anteriores es ordenándolos de **forma acumulativa**; esto es, yendo desde un 100% hasta un 0% (Figura 17). En este caso se aprecia de manera más sencilla que el número de alumnos que comentan decrece conforme avanza el curso. El pico que se producía en la gráfica anterior en torno al 2 de junio coincide con la caída que sufre esta gráfica en torno a la dicha fecha. Como se ha indicado anteriormente, esta fecha fue la última en la que se liberaron materiales nuevos.

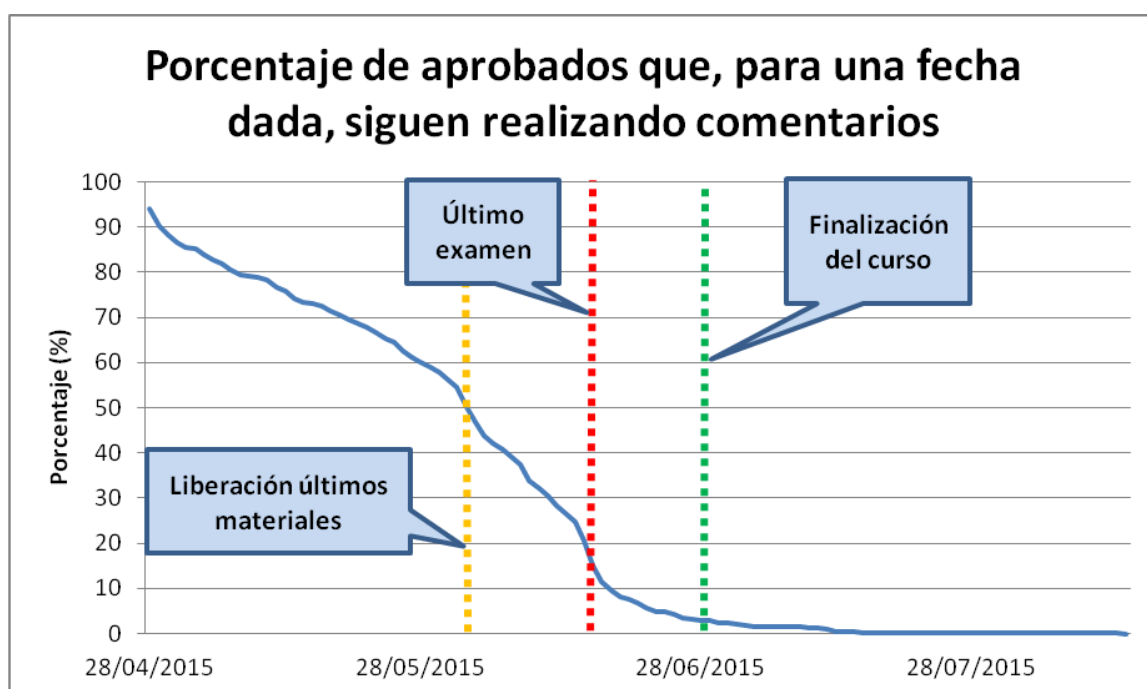


Figura 17. Porcentaje de alumnos aprobados que, para una fecha dada, siguen realizando comentarios.

5.2 Top contributors

A continuación se mostrarán los resultados obtenidos que están más relacionados con los mayores contribuidores. Primero se puede analizar si los *top contributors* tienen tal título por haber **comentado a lo largo de todo el curso** o si, por el contrario, hubo un **pequeño periodo de tiempo en el que realizaron multitud de aportaciones**. En este caso de estudio, entre los *top 10* hay algunas diferencias (Figura 18 y Figura 19). Por ejemplo, considerando que el curso tuvo lugar durante 64 días, el *top 1* realizó aportaciones en 51 de esos 64 días (aunque haya algunos días con más comentarios que otros). En cambio, el usuario *top 4* solamente realizó aportaciones en 15 días. Otra comparación se puede establecer entre los usuarios *top 6* y *top 10*: el primero comentó 24 días y el segundo solamente 16. En la Tabla 5 se pueden ver estos valores para los 20 mayores contribuidores.

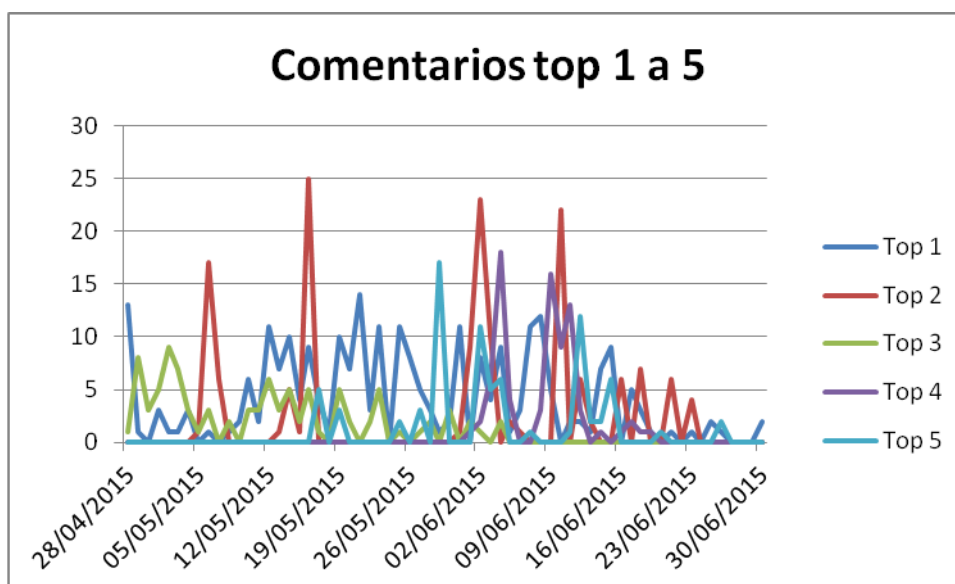


Figura 18. Evolución temporal de los comentarios desde el mayor contribuidor número 1 hasta el mayor contribuidor número 5.

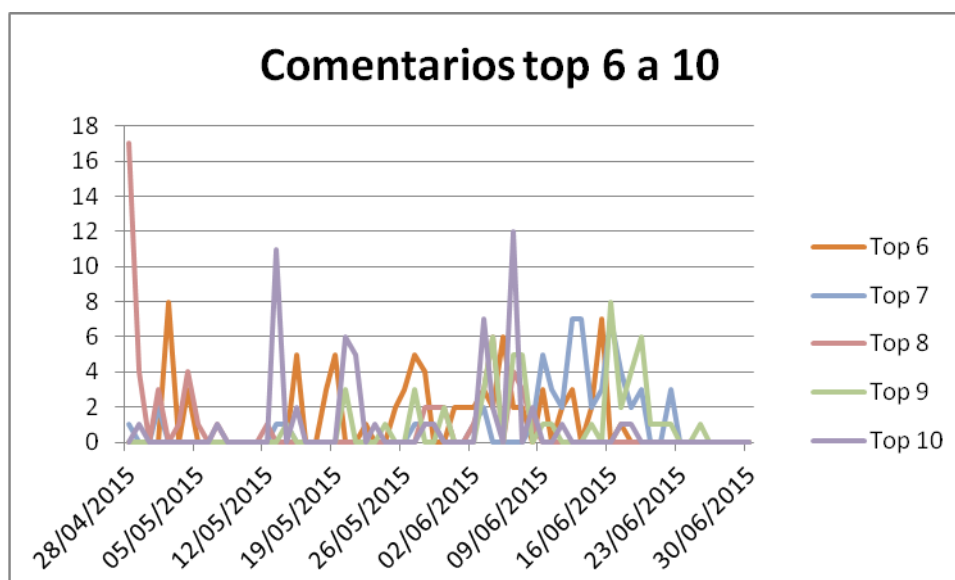


Figura 19. Evolución temporal de los comentarios desde el mayor contribuidor número 6 hasta el mayor contribuidor número 10.

Los datos anteriores correspondían a los 10 mayores contribuidores; se ha ampliado el mismo estudio a los 10 siguientes (Figura 20 y Figura 21). Al igual que ocurría con los 10 primeros mayores contribuidores, en esta ocasión también hay diferentes casos. Por ejemplo, el usuario *top 11* aportó en 29 días mientras que el *top 13* solamente lo hizo en 11 días. Como última comparación, el usuario *top 18* realizó aportaciones al foro en 24 días mientras que el usuario *top 19* solamente aportó en 9 días. De nuevo, en la Tabla 5 se pueden ver estos valores para los 20 mayores contribuidores.

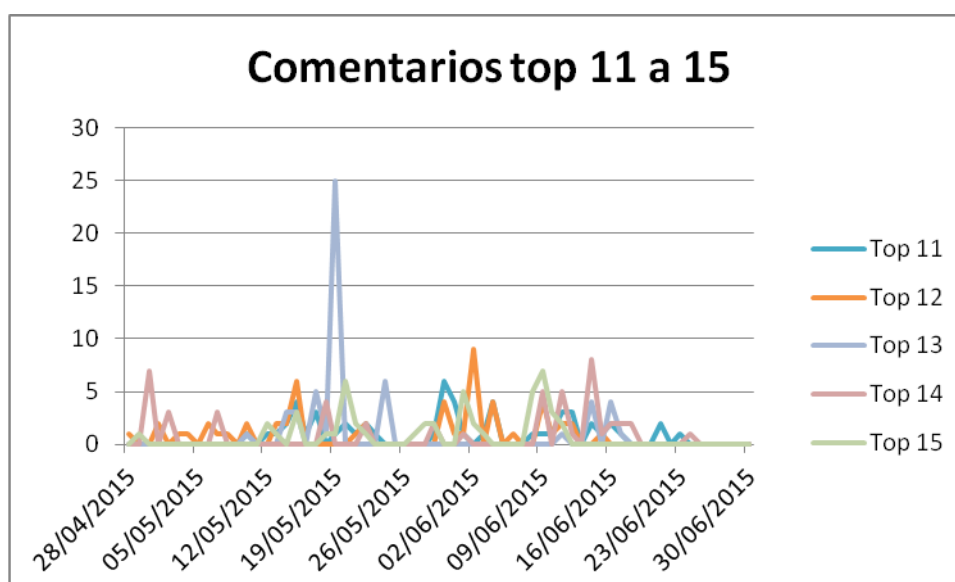


Figura 20. Evolución temporal de los comentarios desde el mayor contribuidor número 11 hasta el mayor contribuidor número 15.

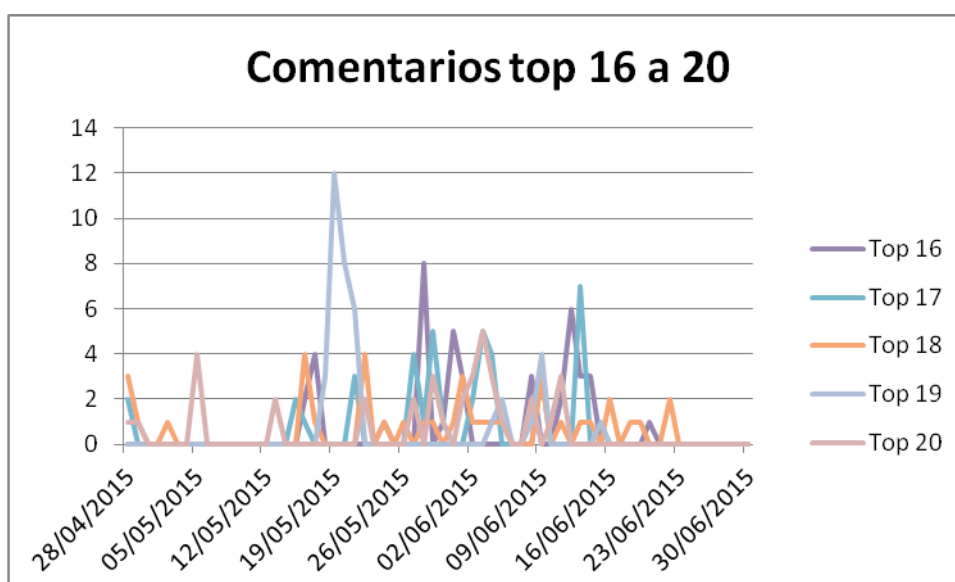


Figura 21. Evolución temporal de los comentarios desde el mayor contribuidor número 16 hasta el mayor contribuidor número 20.

	Aportaciones	Días aportados	Días no aportados	Días aportados (%)	Días no aportados (%)
Top 1	256	51	13	79,7	20,3
Top 2	155	20	44	31,2	68,8
Top 3	96	29	35	45,3	54,7
Top 4	81	15	49	23,4	76,6
Top 5	79	16	48	25	75
Top 6	78	24	40	37,5	62,5
Top 7	60	22	42	34,4	65,6
Top 8	56	17	47	26,6	73,4
Top 9	56	20	44	31,25	68,75
Top 10	55	16	48	25	75
Top 11	55	29	35	45,3	54,7
Top 12	54	24	40	37,5	62,5
Top 13	54	11	53	17,2	82,8
Top 14	50	16	48	25	75
Top 15	48	19	45	29,7	70,3
Top 16	43	13	51	20,3	79,7
Top 17	42	14	50	21,9	78,1
Top 18	38	24	40	37,5	62,5
Top 19	38	9	55	14,1	85,9
Top 20	36	16	48	25	75

Tabla 5. Número de aportaciones al foro y número y porcentaje de días aportados de los 20 mayores contribuidores. Nota: el curso duró 64 días (del 28 de abril al 30 de junio de 2015).

De la tabla anterior resulta interesante que solamente en uno de los 20 usuarios (concretamente, en el *top 1*) el número de días que comentó es mayor que el número de días que no realizó aportaciones. Se realizaron las mismas cuentas para las dos primeras

semanas del curso y los resultados aparecen en la Tabla 6. En este caso, hay 4 usuarios cuyo número de días en los que sí realizaron aportaciones es mayor que el número de días en los que no. También es destacable que en estas dos semanas todavía había 5 *top contributors* que aún no habían realizado ninguna aportación; tal vez sean *latecomers* que se incorporan tarde al curso. De ser así, quedaría resaltada la importancia de este tipo de participantes

	Aportaciones	Días aportados	Días no aportados	Días aportados (%)	Días no aportados (%)
Top 1	45	12	3	80	20
Top 2	24	3	12	20	80
Top 3	54	13	2	86,7	13,3
Top 4	0	0	15	0	100
Top 5	0	0	15	0	100
Top 6	11	2	13	13,3	86,7
Top 7	3	2	13	13,3	86,7
Top 8	32	8	7	53,3	46,7
Top 9	0	0	15	0	100
Top 10	2	2	13	13,3	86,7
Top 11	2	2	13	13,3	86,7
Top 12	11	8	7	53,3	46,7
Top 13	1	1	14	6,7	93,3
Top 14	13	3	12	20	80
Top 15	3	2	13	13,3	86,7
Top 16	0	0	15	0	100
Top 17	2	1	14	6,7	93,3
Top 18	5	3	12	20	80
Top 19	0	0	15	0	100
Top 20	6	3	12	20	80

Tabla 6. Número de aportaciones al foro y número y porcentaje de días aportados de los 20 mayores contribuidores durante las dos primeras semanas del curso.

Otra cuestión que cabe plantearse es la de conocer si los *top contributors* son **creadores de respuestas o iniciadores de discusión**. Se realizó la investigación para los *top 10* y los resultados están apuntados en la Tabla 7. Aunque en los 10 casos el número de respuestas es mayor que el de inicios, los porcentajes más cercanos se dan en el usuario *top 8* (55,4% respuestas y 44,6% inicios). Efectivamente, inspeccionando un poco los hilos creados por éste, la mayoría de ellos iban destinados al *staff* del *MOOC* informando sobre errores ortográficos, fallos en algún ejercicio o solicitando información sobre fechas de publicación de ciertos materiales.

	Respuestas	Inicios	Respuestas (%)	Inicios (%)
Top 1	238	18	93	7
Top 2	124	31	80	20
Top 3	92	4	95,8	4,2
Top 4	77	4	95,1	4,9
Top 5	75	4	94,9	5,1
Top 6	52	26	66,7	33,3
Top 7	46	14	76,7	23,3
Top 8	31	25	55,4	44,6
Top 9	50	6	89,3	10,7
Top 10	41	14	74,5	25,5

Tabla 7. Número y porcentaje de inicios de discusión y respuestas a discusiones ya creadas de los 10 mayores contribuidores.

Otro aspecto interesante por conocer engloba a las **relaciones y comunidades** que han podido surgir entre los usuarios del foro. Esto es, conocer si algún usuario responde a cualquier alumno o si solamente se comunica con un grupo reducido de ellos. Para analizar este aspecto se han utilizado gráficos sociales. El primero que se presenta corresponde a las relaciones establecidas por todos los alumnos (Figura 22). Cada punto representa a un usuario que participó en el foro y las uniones representan enlaces entre usuarios, los cuales se producen al responder o ser respondido. Aunque no es un gráfico muy esclarecedor dada la alta densidad de información, se pueden destacar alguna característica. Por ejemplo, se identifica bien quiénes son los dos mayores contribuidores. Los puntos que no están unidos con ningún otro son usuarios que publicaron comentarios solamente en publicaciones creadas por ellos mismo sin responder a nadie concretamente.

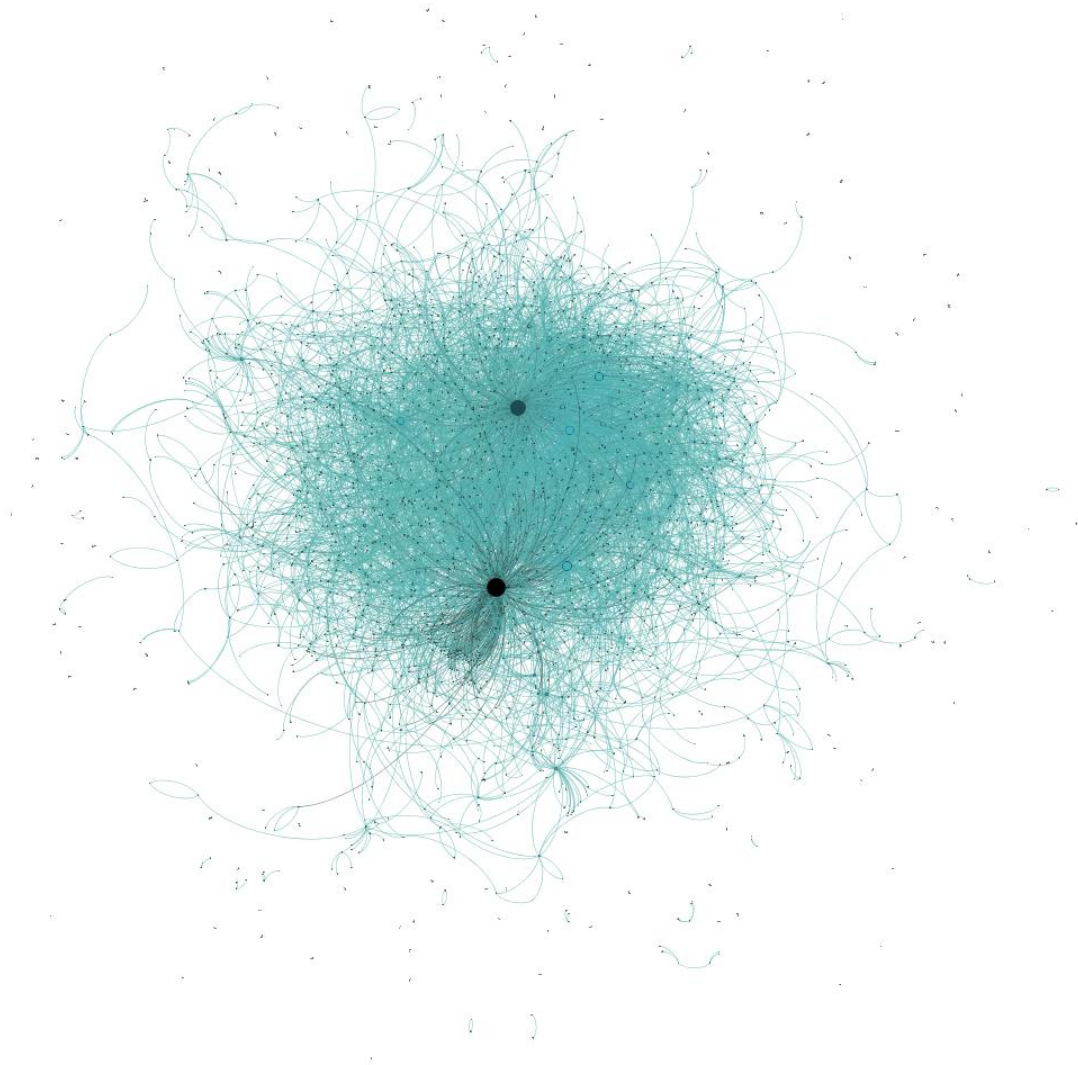


Figura 22. Diagrama social de todos los alumnos del curso.

Un gráfico similar fue obtenido considerando como “respondedores” solamente a los 20 mayores contribuidores (Figura 23). De nuevo, los puntos son usuarios y los enlaces son uniones entre ellos. Aunque en la figura anterior se han oscurecido los colores para que pudieran visualizarse mejor las líneas más claras, el significado de los enlaces y de los puntos cambia según el color que posean. Así, un punto o nodo con un tono más oscuro será un participante que habrá respondido a más usuarios, y el color del enlace o arista va en concordancia con esta característica. Cabe notar que no es posible visualizar en este gráfico las veces que un usuario ha contestado a otro (esto es, una unión entre dos nodos significa que ha habido relación entre ellos al menos una vez, pero no se aporta información de cuántas veces en concreto).

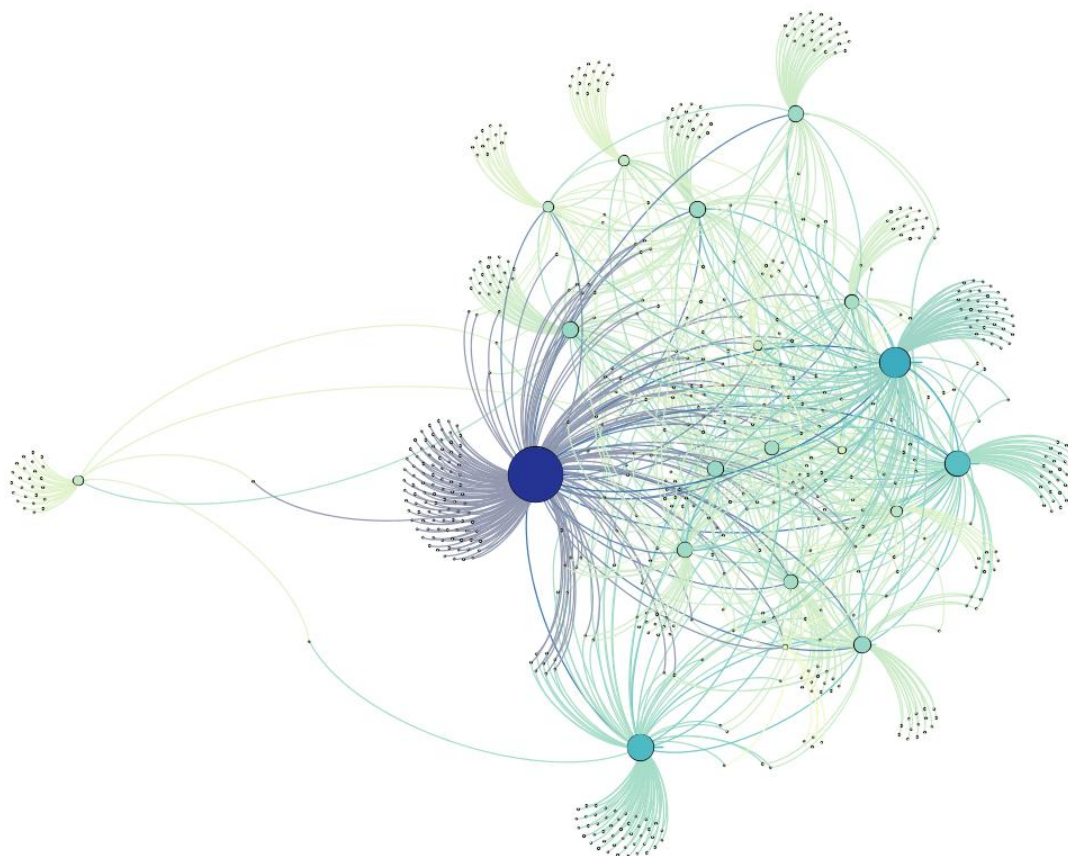


Figura 23. Diagrama social considerando como “respondedores” a los 20 *top contributors*.

Finalmente, también se realizó una gráfica para los 10 mayores contribuidores (Figura 24). Como también se podía observar en el diagrama anterior, estos mayores contribuidores responden a muchos usuarios distintos e incluso llegan a relacionarse entre ellos mismos.

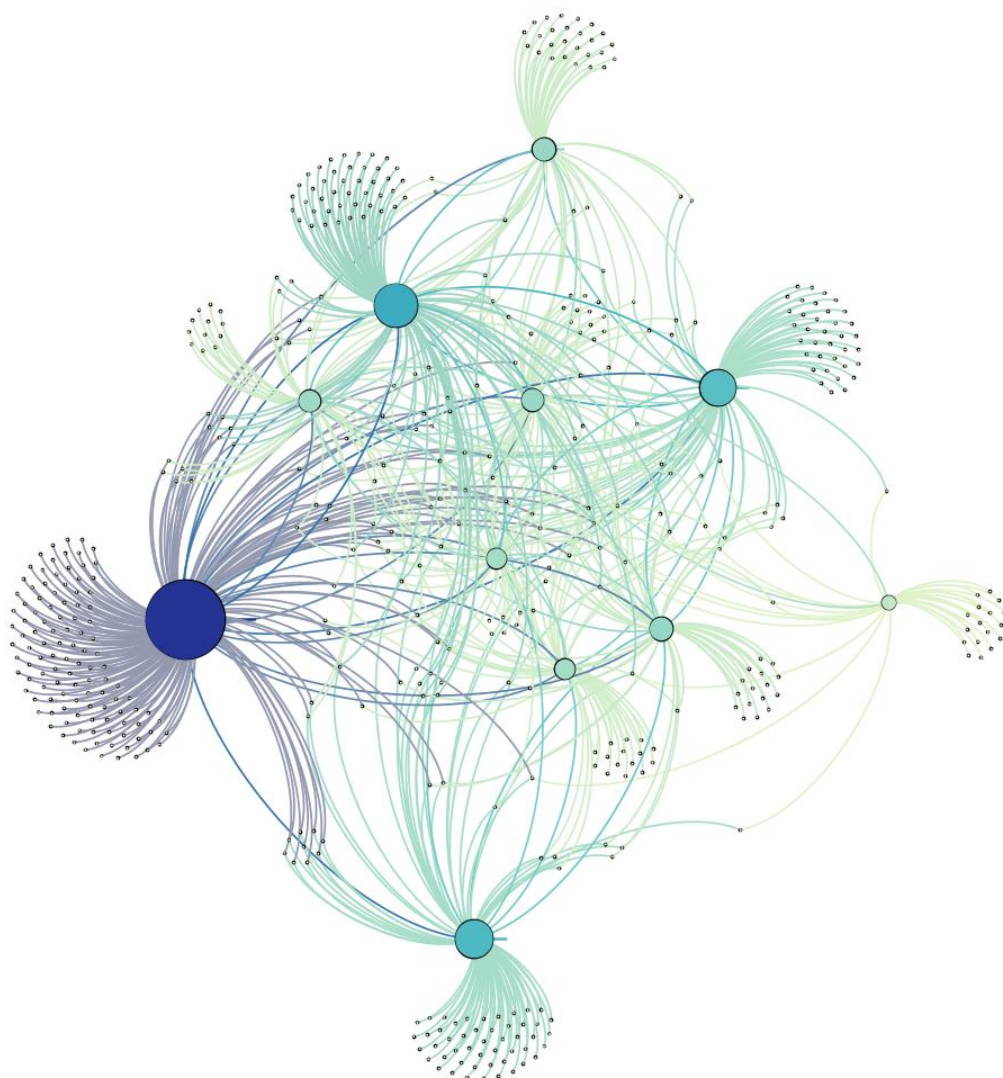


Figura 24. Diagrama social considerando como “respondedores” a los 10 *top contributors*.

Chapter 6³

Conclusions and future work

6.1 Conclusions

Before starting with the conclusions, the **primary objectives** will be reminded: on one side, checking whether the hypotheses and conclusions stated by other scientific papers hold for this case of study; on the other side, studying the top contributors' characteristics and trying to predict early which users will become these more active participants. As the results were organized firstly with general course characteristics and secondly with top contributor's features, the same classification will be followed in this chapter. It is also reminded that the results shown in this section have been obtained by studying a MOOC titled "Introduction to Programming with Java Part 1: Starting to Program with Java".

First of all, the most relevant aspects of the use of the social network *Twitter* will be showed. Clearly, the discussion forum of the course has been more widely used and has had more impact than the "tweets". In one paper from the related work [13] it was already noted that the use of *Twitter* was lesser than the use of the forum in the MOOC "Digital Education of the Future". Regarding the topic of these "tweets" (Table 3), 47.7% of these were about communicating and spreading the course. Something similar happened in the just-mentioned course, even though in that case the percentage was lesser since some messages were unrelated to the MOOC, what has not happened in this case. Because of this, it is concluded that *Twitter* is more useful as a tool for the course to become known

³ Para la traducción en español de este capítulo, consultar el Anexo IV.

than as a source for links between students to appear; in any case, it is advisable to keep checking these results with the ones from other courses.

As a **general** course feature, only 2% of registered participants passed the MOOC (Figure 3). This fulfils the premise stated in Chapter 2 in which it was said that the number of students that pass a MOOC is typically lower than 10% of the ones that register in it. Making a comparison, in the MOOC “Digital Education of the Future” [13] 8% of registered users passed the course (456 students). Even though these percentages and values are different since the number of enrolled users in each MOOC are far from each other, in both cases less than 10% of participants passed the course, as it was just stated.

In this course, 6% of all participants commented in the forum and the remaining 94% did not (Figure 3). It was computed that 95% of the users that did not comment on the forum failed the course (Figure 5). This goes in accordance with the fact that the vast majority of registered users do not participate in the course (neither in the discussions nor in the activities). Regarding the remaining 5% of students who failed, most of these will probably be users that greeted in the welcoming discussions but then dropped out the course, which appears to be confirmed by Figure 15. On the other side, with respect to the participants that passed, it is interesting that half of them contributed to the forum and the other half did not (Figure 4). This is significantly noteworthy since a highly positive correlation between passing and commenting could be expected; this is, that it existed a greater number of students that passed the course and commented than the number of users that passed and did not comment. Still, if all data is analyzed as a whole (Figure 3), according to the information obtained in this case of study, the following predictions can be done: if a user comments in the forum, there is a 17% chance that the student passes the course and a 83% chance that the user fails; if a participant does not contribute to the forum, there is a 1% chance that he or she passes and a 99% chance that he or she fails. Expressing these conditional probabilities in mathematical form:

$$P(\text{pass}|\text{comment}) = \frac{1}{2} = 50\% \qquad P(\text{comment}|\text{pass}) = \frac{1}{6} \approx 17\%$$

With respect to learners’ final marks (Figure 6), as it was mentioned in Chapter 5, it can be deduced that most participants that do not obtain a zero as final mark only take one evaluation activity, probably the first one, before disengaging from the course. Also, the correlation between final mark and number of contributions was computed, obtaining the value 0.21. Even though it is a positive value, it is also low, and therefore further statistical analyses are required in order to obtain meaningful conclusions. The same happens when using a relation between created threads and comments published in already created discussion, whose correlation with final marks is 0.25.

Continuing with conclusions, the graph that represents the evolution of the number of comments according to the passing of time (Figure 7) corresponds to what usually happens in these MOOCs (again, due to the excitement produced at the beginning of the course and the welcoming discussions). In this line, there is a large portion of the messages during the first days and then a pronounced drop takes place. Later, some peaks are reached, especially around days in which new material is provided. In agreement with this aspect and with the previous passing and failing classification, it is known that most participants (3699, 68% of all) only contributed once (Figure 8). The number of users that

commented twice is reduced to 689 (13% of all). The mean is 2.5 messages per user (taking only into account the students that contributed).

With respect to the time periods in which the number of messages is greater, it should obviously be taken into account the time difference among distinct zones of the world (Figure 11). In this study it was found that United States and Europe's evenings have more weight than India's one (these locations have been selected as references since most of the users belong to these places).

As stated before, that data from this case of study indicates that 6% of registered users made at least one contribution to the forum, even though only 1% of all both commented and passed the MOOC. In line with this aspect, it was calculated that, the more contributions a user has made, the greater the probability that this user has passed (Figure 13). Intuitively, this corresponds with the students' engagement in the MOOC. For instance, in participants that made at least 8 contributions, the uncertainty of passing or failing is at 50%. From 3 comments on, the graph can be approximated as a line (until reaching 100% of passing users). Something similar happens if only the first two weeks of the course are taken into account, although the approximation to a line is less reliable. Regarding data from the entire course, it can be easily obtained the equation of the aforementioned line (reminding that it should not be applied for percentages lesser than 30%):

$$\left. \begin{array}{l} P(\text{pass} | \text{comment} > 9 \text{ times}) = 50\% \\ P(\text{pass} | \text{comentar} > 24 \text{ times}) = 75\% \\ P(\text{pass} | \text{comment} > 36 \text{ times}) = 100\% \end{array} \right\} \rightarrow \% \text{Passing} \approx 1,9 \cdot \text{NumberComments} + 32,9$$

With respect to the students that both passed the course and commented in the forum, information relating their last comment date was computed (Figure 16). On the last day available to take the last evaluation activity, around 20% of users that commented and passed kept contributing (Figure 17). It could be underlined that, even though the number of participants contributing decreases as the course goes by, there exists a notable drop around the 2nd of June, last date in which new material was provided. Thus, it is possible that some users have taken the evaluation activity related to this material shortly after it was made available, and then they have not checked the forums anymore since the course was over for them.

Before providing specific data about top contributors, it will be reminded that the early identification of such contributors is related to awarding them with special moderator permissions in the discussion forums (the platform *edX* calls them "Community Teaching Assistant" or "Community T.A."). Particularizing obtained information on **top contributors**, 17 of them contributed at least 25% of the days the course lasted. Some of them commented more frequently and some others contributed sporadically. In addition, the early identification of these users is a complex task according to the obtained data. For instance, 5 of 20 top contributors did not comment during the first two weeks of the course (maybe they were latecomers, what would strengthen the importance of this group as stated in some papers [13]), and another 6 of them made less than 6 contributions (Table 6). If these top contributors are sorted in descendent number of comments taking into account only the first two weeks, 8 of the "true" top contributors of all MOOC can be found among them (specifically, tops 1, 2, 3,

6, 8, 12, 13 and 20). This way, if early in the course the top 10 were selected, among them there would be 5 of the “true” top contributors.

Regarding their contributions to the forum, the top 10 participants have posted messages in already created discussions more than they have created new threads (Table 7); only in 3 of them the percentage of responses is lesser than 75% and the percentage of started discussions greater than 25%. In one case, the percentage of started threads is that of 44.6% of total published messages; this user usually created threads in which he reported errors to the staff or asked for information about the publication dates of some material.

With respect to links formed by the top contributors with other users (Figure 23), it can be concluded that these contributors do not seem to focus in just a few users. Thus, even though there are participants that are only connected to one top contributor, many of them have connections with more than one. Even there exists links among the top contributors themselves.

6.2 Acquired competences

As student in charge of this project, I have acquired abilities and competences during the development of this work, like happens in any other course.

The most relevant concept may be the difference the Bachelor’s thesis has with respect to the rest of Bachelor courses, since, if I am not mistaken, this has been the first time in which I have faced an issue without a teacher showing me new concepts each week (even though several meetings with the coordinators have taken place). Along this line, for the first time the issue at hand did not have a solution that could be looked up in a solution book, for example. Indeed, this has been a project with a research component since we did not know what the results would be. In a certain way, something similar happens in other courses when we are requested to do an assignment or project, but the difference resides in the fact that, in those cases, as the course progresses, we are taught new concepts with which complete said task. Besides, I have had to read different scientific papers as initial work, many of them written in English. This aspect, as well as serving as starting point and checking which experiments were already carried out, has provided me with general knowledge about MOOCs and has increased my understanding about their characteristics and operation.

As a complement of what has been said, again this has been the first time I have faced a data organization and classification issue. For that, apart from learning some properties I did not know about *Microsoft Excel*, I have used algorithms programmed in *MATLAB*, a tool I was presented during my first year at University and that we have used in quite some courses. In this way, I really consider that I have tested my programming knowledge. As a slight curiosity I will state that all issue around Programming fascinated me since I had my first subject about it the first year at University, and therefore I have faced this challenge with pleasure and initiative.

All things considered, and ending this section, I believe I have obtained satisfactory results with regard to contributing to my training as a student and as a person.

6.3 Future work

As future work, the main line that can be followed is the **study of top contributors** in different ways. Inside this very same MOOC, additional studies can be done, like inspecting how technical the contributions are. In this line, it could be checked if each one comments only about one issue or not. Another aspect that was not taken into account in this work was top contributors' profiles. Thus, if more personal information about these contributors was studied, there could be some common features among them (above all, it may be relevant taking into consideration their previous training or studies).

In more technical lines and aspects whose automation would be simple, new algorithms could be programmed. It may be interesting to know the time periods in which top contributors comment (that is, obtaining a graph similar to Figure 11 taking into account only top contributors). The contributions that have more positive votes could be selected, and then it could be checked to which contributors they belong. In addition, along this line, the field “endorsed” could also be taken into account for a similar analysis (it is reminded that, when this field is followed by “true”, a moderator or teacher as marked a message as a valuable contribution). With respect to more complex statistical studies, more correlation analyses could be made by selecting only the number of contributions and the marks of each week; that is, checking the marks of the activities of each week and relating them to the number of messages associated to that week.

Once several studies about identifying top contributors have been carried out, and the amount of information is large enough to make a train group and a test group, analyses like linear regressions and descriptors extraction could be made. In this sense, an algorithm that adequately classifies the train group and then evaluates the test group would be programmed. If no satisfactory results are obtained, the code is modified and the program is executed again. It could be interesting to have a very high detection probability in exchange of having a quite high false alarm probability (that is, detecting practically all top contributors but selecting also some users that do not belong to this group). Besides, the results could be refined by making similar analyses each week and by considering all accumulated course data.

Outside this MOOC, it could be analyzed how the users categorized here as top contributors behave in other courses. Also, it could be checked if these users actively use social networks like *Twitter* in their daily life. Of course, it could be studied whether the data obtained in this project resemble other courses or not. It would be interesting to divide these courses between the ones with topics that are similar to the one this work studies and courses with different subjects (this is, it may happen that the interactions among users changes drastically in a MOOC about Programming and a MOOC about Medicine).

Referencias

1. Alario Hoyos, C.; Pérez Sanagustín, M.; Delgado Kloos, C.; Parada G., H. A.; Muñoz Organero, M. Noviembre de 2014. *Who are the top contributors in a MOOC? Relating participants' performance and contributions*. Journal of Computer Assisted Learning. Volumen 32, número 3, páginas 232 a 243. (Disponible en: <http://onlinelibrary.wiley.com/doi/10.1111/jcal.12127/abstract>. Consultado el: 20 de octubre de 2015).
2. Hew, K. F.; Cheung, W. S. 27 de mayo de 2014. *Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges*. (Disponible en: <http://www.sciencedirect.com/science/article/pii/S1747938X14000128>. Consultado el: 20 de octubre de 2015).
3. Junio de 2013. *SCOPEO INFORME N°2: MOOC: Estado de la situación actual, posibilidades, retos y futuro*. Número 2. (Disponible en: <http://scopeo.usal.es/wp-content/uploads/2013/06/scopeoi002.pdf>. Consultado el: 7 de junio de 2016).
4. MiríadaX. (Disponible en: <https://miriadax.net/home>. Consultado el: 29 de mayo de 2016).
5. edX. (Disponible en: <https://www.edx.org/>. Consultado el: 29 de mayo de 2016).
6. Shah, D. 14 de diciembre de 2015. *MOOC Trends in 2015: The Death of Free Certificates; How Much Do Different MOOC Providers Charge for a Certificate in 2015?* Class Central. (Disponible en: <https://www.class-central.com/report/death-of-free-certificates/>. Consultado el: 30 de mayo de 2016).
7. Pappano, L. 2 de noviembre de 2012. *The Year of the MOOC*. The New York Times. (Disponible en: http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?pagewanted=all&_r=1. Consultado el: 29 de mayo de 2016).
8. Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal. Boletín Oficial del Estado. (Disponible en: <https://www.boe.es/buscar/pdf/1999/BOE-A-1999-23750-consolidado.pdf>. Consultado el 1 de junio de 2016).

BIBLIOGRAFÍA: REFERENCIAS

9. *Guía de Seguridad de Datos*. Agencia Española de Protección de Datos. (Disponible en: http://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/comm on/Guias/GUIA_SEGURIDAD_2010.pdf. Consultado el: 6 de junio de 2016).
10. Terms of Service. edX. (Disponible en: <https://www.edx.org/edx-terms-service>. Consultado el: 30 de mayo de 2016).
11. Privacy Policy. edX. (Disponible en: <https://www.edx.org/edx-privacy-policy>. Consultado el: 30 de mayo de 2016).
12. Ho, A. D.; Reich, J.; Nesterko, S.; Seaton, D. T.; Mullaney, T.; Waldo, J.; Chuang, I. 21 de enero de 2014. *HarvardX and MITx: The First Year of Open Online Courses*. HarvardX and MITx Working Paper No. 1. (Disponible en: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263. Consultado el: 20 de octubre de 2015).
13. Alario Hoyos, C.; Pérez Sanagustín, M.; Delgado Kloos, C.; Parada G., H. A.; Muñoz Organero, M. 2014. *Delving into Participants' Profiles and Use of Social Tools in MOOCs*. IEEE Transactions on Learning Technologies. Volumen 7, número 3, páginas 260 a 266. (Disponible en: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6766737>. Consultado el: 20 de octubre de 2015).
14. 10 de abril de 2014. *La educación digital; Los MOOC (II)*. Fundación España Digital. (Disponible en: <http://www.espanadigital.org/la-educacion-digital-los-mooc-ii/>. Consultado el: 7 de junio de 2016).
15. *Estado del arte y oportunidades de negocio. Tecnología digital aplicada al entorno formativo*. Escuela de Organización Industrial, Ayuntamiento de Málaga, Fondo Europeo de Desarrollo Regional. (Disponible en: http://www.digitalmalaga.com/fileInforme/13/Estado%20arte_oportunidades_Tecnolog%C3%ADa%20aplicada%20al%20Entorno%20Formativo%20v1.0%20def.pdf. Consultado el: 7 de junio de 2016).
16. Siemens, G. 12 de diciembre de 2004. *Connectivism: A Learning Theory for the Digital Age*. (Disponible en: <http://web.archive.org/web/20140527130019/http://www.elearnspace.org/Articles/connectivism.htm>. Consultado el: 7 de junio de 2016).
17. Downes, S. 9 de abril de 2013. *What the 'x' in 'xMOOC' stands for*. (Disponible en: <https://plus.google.com/+StephenDownes/posts/LEwaKxL2MaM>. Consultado el: 7 de junio de 2016).
18. 22 de enero de 2016. *cMooc vs. xMooc*. Mooc.es. (Disponible en: <http://mooc.es/cmoox-vs-xmooc/>. Consultado el: 7 de junio de 2016).
19. *El conductismo*. Enfoques educativos. (Disponible en: <http://hadoc.azc.uam.mx/enfoques/conductismo.htm>. Consultado el: 7 de junio de 2016).
20. *Flipped classroom*. Wikipedia. (Disponible en: https://en.wikipedia.org/wiki/Flipped_classroom. Consultado el: 7 de junio de 2016).
21. *Visión – What is the Flipped Classroom*. The flipped classroom. (Disponible en: <http://www.theflippedclassroom.es/what-is-innovacion-educativa/>. Consultado el: 7 de junio de 2016).
22. Álvarez, D. 15 de octubre de 2012. *¿Le damos la vuelta al aula...? The Flipped Classroom*. e-aprendizaje. (Disponible en: <http://e-aprendizaje.es/2012/10/15/le-damos-la-vuelta-al-aula-the-flipped-classroom/>. Consultado el: 7 de junio de 2016).

23. *What is NoSQL?* MongoDB. (Disponible en: <https://www.mongodb.com/nosql-explained>. Consultado el: 5 de junio de 2016).
24. *NoSQL*. Wikipedia. (Disponible en: <https://es.wikipedia.org/wiki/NoSQL>. Consultado el: 5 de junio de 2016).
25. Microsoft Excel. (Disponible en: <https://products.office.com/es-es/excel>. Consultado el: 5 de junio de 2016).
26. MATLAB. (Disponible en: <http://es.mathworks.com/products/matlab/>. Consultado el: 5 de junio de 2016).
27. NodeXL. (Disponible en: <http://nodexl.codeplex.com/>. Consultado el: 5 de junio de 2016).
28. Gephi. (Disponible en: <https://gephi.org/>. Consultado el: 5 de junio de 2016).
29. Microsoft Word. (Disponible en: <https://products.office.com/es-es/word>. Consultado el: 5 de junio de 2016).
30. ProjectLibre. (Disponible en: <http://www.projectlibre.org/>. Consultado el: 15 de junio de 2016).
31. *Introduction to Programming with Java Part 1: Starting to Code with Java*. edX. (Disponible en: <https://www.edx.org/course/introduction-programming-java-part-1-uc3mx-it-1-1x-0>. Consultado el: 26 de mayo de 2016).
32. *Introduction to Programming with Java Part 2: Writing Good Code*. edX. (Disponible en: <https://www.edx.org/course/introduction-programming-java-part-2-uc3mx-it-1-2x>. Consultado el: 12 de junio de 2016).
33. Hoi K. Suen. Julio de 2014. *Peer Assessment for Massive Open Online Courses (MOOCs)*. The International Review of Research in Open and Distributed Learning. Volumen 15, número 3. (Disponible en: <http://www.irrodl.org/index.php/irrodl/article/view/1680/2904>. Consultado el: 13 de junio de 2016).
34. Ztreamy. (Disponible en: <http://www.ztreamy.org/>. Consultado el: 24 de mayo de 2016).
35. Discussion Forums Data. edX Research Guide. (Disponible en: https://devdata.readthedocs.io/en/latest/internal_data_formats/discussion_data.html. Consultado el: 23 de mayo de 2016).
36. Worldtimeserver. (Disponible en: http://www.worldtimeserver.com/current_time_in.UTC.aspx. Consultado el: 12 de junio de 2016).
37. MATLAB Student. MathWorks. (Disponible en: https://es.mathworks.com/academia/student_version/. Consultado el: 5 de junio de 2016).
38. Frequently Asked Questions. MathWorks. (Disponible en: <https://es.mathworks.com/store/link/faq>. Consultado el: 5 de junio de 2016).

BIBLIOGRAFÍA: REFERENCIAS

Anexo I

Planificación

Este proyecto comenzó en septiembre de 2015 y finalizó en junio de 2016. En este Trabajo de Fin de Grado, al haberse tratado de un proyecto con un grado de investigación, no ha habido una estructura lineal en cuanto a forma de proceder. En cambio, se ha utilizado un **proceso iterativo** esquematizado en la Figura 25. Explicando dicho proceso, como ya se comentó en el Capítulo 2, se comenzó por la lectura de artículos proporcionados por los coordinadores y de otros buscados por el propio alumno. Algunos de estos documentos exponían rasgos generales de los *MOOCs* y otros utilizaban como caso de estudio algún curso en concreto sobre el que explicaban características y conclusiones. Tras la lectura de dichos documentos, se elaboró un breve informe en el que se recogían los aspectos más destacables de esos artículos y los puntos de partida por los que se podría comenzar en este proyecto. A continuación, adjudicado ya el Trabajo de Fin de Grado y disponiendo de los *tweets*, empezó el proceso iterativo que se mencionó anteriormente. Dicho proceso comenzaba con la recogida de ideas tanto por parte de los coordinadores como del alumno y la proposición de nuevas líneas sobre las que trabajar. Después, se examinaban los datos de los que se disponían y a menudo se reorganizaban y manipulaban los mismos para, por ejemplo, obtener una lectura óptima con *MATLAB*. El siguiente paso consistía en la creación de algoritmos, en caso de usar dicho *software*, o la utilización de *Excel* o *Gephi*, según fuera el estudio a realizar. Tras esto, se guardaban, clasificaban y organizaban los resultados obtenidos y se analizaban de forma conjunta entre el alumno y los coordinadores. Y finalmente, el último paso antes

ANEXO I: PLANIFICACIÓN

de volver al primer punto de este sistema iterativo trataba de seleccionar qué resultados se consideraban satisfactorios e ir actualizando la memoria final del proyecto.

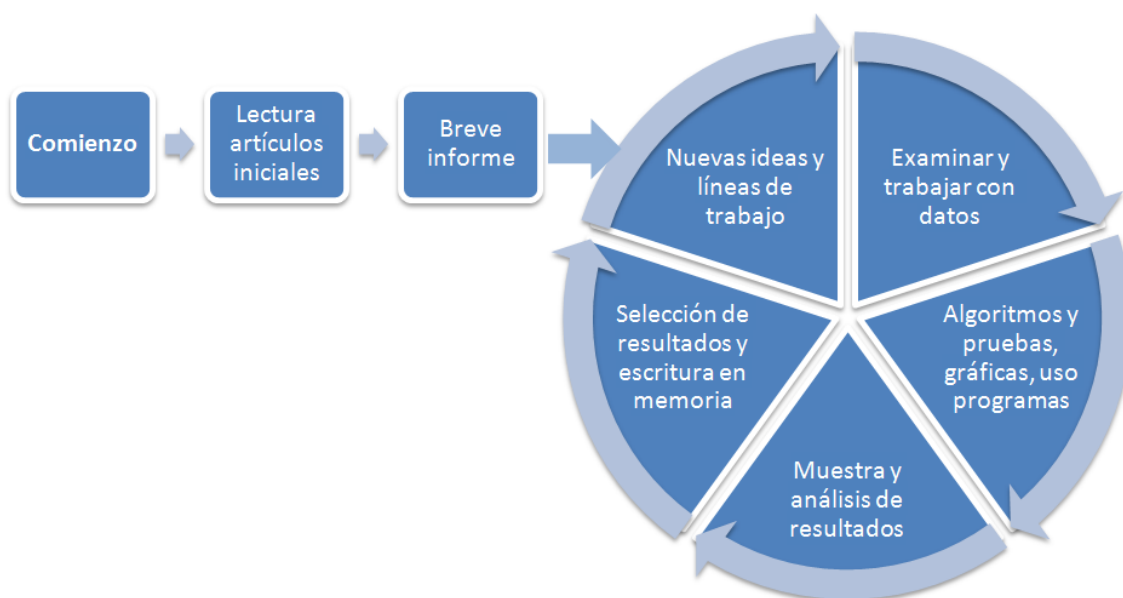


Figura 25. Esquema de la planificación del proyecto.

Como se ha dicho que no se ha seguido un proceso en cascada, un diagrama de *Gantt* no es óptimo para la representación de ese caso. Aún así, y de acuerdo con la Figura 25, un acercamiento que se puede realizar es el indicado en la Figura 26. Así, la mayor parte del tiempo fue ocupada por el proceso iterativo. Se realizaron aproximadamente siete iteraciones y el paso de la fase “aceptación de resultados y memoria” a la fase “nuevas ideas” solía tener lugar durante las reuniones presenciales entre el alumno y los coordinadores.



Figura 26. Diagrama de *Gantt* del proceso de organización del proyecto. Nota: no tener en cuenta las horas; fijarse solamente en los días y entenderlos como días completos.

En cuanto a las horas de trabajo (Tabla 8), como este proyecto ha sido realizado en paralelo a todas las demás asignaturas de 4^º curso, hubo mayor actividad en periodos en los que no se tenía mucha carga por parte de estas asignaturas (enero y febrero tras los exámenes del primer cuatrimestre y antes de que las asignaturas entraran en profundidad con la materia, y finales de mayo y junio por la finalización de los exámenes del segundo cuatrimestre). Así, se han estimado 291 horas de trabajo por parte del alumno.

<u>Mes</u>	<u>Horas</u>	<u>Tareas principales</u>
Septiembre	18	Artículos iniciales y elaboración informe
Octubre	19	Artículos adicionales y <i>tweets</i>
Noviembre	20	Tweets y fichero desechado
Diciembre	20	Trabajo con ficheros proporcionado por <i>edX</i>
Enero	26	Trabajo con ficheros proporcionado por <i>edX</i>
Febrero	26	Trabajo con ficheros proporcionado por <i>edX</i>
Marzo	17	Organización y cálculos <i>Excel</i> , código <i>MATLAB</i>
Abril	17	Organización <i>Excel</i> , código <i>MATLAB</i> , gráficas
Mayo	28	Código y memoria
Junio	100	Código y memoria
	291	

Tabla 8. Horas dedicadas a este Trabajo de Fin de Grado⁴.

⁴ Antes de disponer del archivo proporcionado por *edX*, se trabajó con otro fichero que también contenía las contribuciones al foro. Éste se desechó cuando se obtuvo el de la propia plataforma.

ANEXO I: PLANIFICACIÓN

Anexo II

Presupuesto

En cuanto a los programas utilizados, el único que no era de acceso libre fue *MATLAB*. Éste fue el utilizado para diseñar algoritmos con los que obtuvieron datos y algunas gráficas, como ya se ha comentado. La versión de este *software* para estudiantes cuesta 35€ [37] y está disponible indefinidamente [38], por lo que no se calculará ninguna amortización. En línea con esto, se utilizó un ordenador durante los aproximadamente 9 meses que ha durado el proyecto (desde mediados de septiembre hasta mediados de junio). Si se estima una vida útil media de 5 años (60 meses) antes de cambiar ninguna pieza y un coste medio de 600€, la amortización para el periodo de este proyecto es de 90€. Según las tablas de la Universidad Carlos III de Madrid y atendiendo a los cargos de cada participante y a las horas estimadas, se puede realizar una estimación como la reflejada en la Tabla 9.

ANEXO II: PRESUPUESTO

<u>Nombre</u>	<u>Categoría</u>	<u>Horas</u>	<u>Coste/hora</u>	<u>Total €</u>
Carlos Alario Hoyos	Formación Postdoctoral MINECO	33	27,33	901,89
Iria Manuela Estévez Ayres	Titular Universidad Interino	21	31,49	661,29
Javier Velo Beascoechea	Recién Titulado	291	15,27	4443,57
		345		6006,75

Tabla 9. Estimación del presupuesto.

De esta forma, y añadiendo los costes de *MATLAB* y del ordenador, se estima que el presupuesto total de este proyecto asciende a la cantidad de 6131,75€.

Anexo III

Introducción y objetivos

III.1 Introducción

Los *Massive Open Online Courses* (*MOOCs* por sus siglas en inglés, “Cursos Masivos en Línea y Abiertos” en español) han captado la atención de muchas organizaciones educativas superiores. Sus defensores argumentan que los *MOOCs* hacen la educación más accesible al mayor número de personas posible, incrementan el alcance o la visibilidad de la institución que los crea y permiten a los profesores experimentar la ciencia de la enseñanza *online* a un número grande y diverso de estudiantes [1]. Por otra parte, los que se oponen, critican a los *MOOCs* como potencialmente dañinos y los tachan de tecnología alteradora de la educación que ofrece una versión diluida de la misma, al mismo tiempo que incrementan el riesgo de reducciones en el presupuesto de centros de enseñanza públicos [2]. Aunque en este documento no se entrará en esta discusión, se debe reconocer que los *MOOCs* han supuesto un impacto considerable en la forma de enseñar y aprender (aunque probablemente no sustituyan, al menos por el momento, al método tradicional “cara a cara”). De hecho, en uno de los documentos consultados [3] se formula una reflexión sobre si los *MOOCs* sustituirán a la formación universitaria. En dicho documento se considera a los *MOOCs* no como un sustituto de la Educación Superior sino como a un complemento de ésta. Además, ejemplifica su punto de vista realizando una comparación entre un ordenador y una *tablet*: uno no es sustituto del otro; ambos cumplen funciones diferentes. Así, de la misma manera que los *MOOCs*

tienen virtudes que no poseen otros tipos de enseñanza, la formación presencial posee características de las que carecen los *MOOCs*.

Explicando qué son los *MOOCs* en **rasgos generales** [4] [5], se tratan de unos cursos ofertados a través de Internet, desplegados en plataformas como *MiríadaX* o *edX*. La lista de cursos ofertados y la información sobre éstos suele estar disponible para cualquier usuario que navegue por la red, aunque dicho usuario deberá registrarse en la plataforma en cuestión para matricularse en dichos cursos. Al seleccionar un *MOOC*, se suele mostrar información adicional como su duración y su fecha de inicio, entre otros. También suelen estar disponibles una descripción del curso y valoraciones del mismo si no es la primera vez que se oferta, e incluso, en ocasiones, un vídeo presentación del mismo. Una vez inscrito en un curso y cuando éste haya comenzado, irán estando disponibles diferentes materiales otorgados por los coordinadores de dicho curso, tales como vídeos explicativos, documentos escritos sobre los conceptos y tareas para evaluación del alumno. Éste podrá ir completando las tareas a su ritmo siempre y cuando sean realizadas antes de sus fechas de expiración. Dichas tareas pueden ser divididas en dos grupos. Por una parte, ver vídeos y otros contenidos que posean información sobre la materia. Por otra parte están los ejercicios, los cuales en ocasiones se dividen, a su vez, en dos subgrupos. Por un lado, preguntas cerradas, por ejemplo tipo test o multi-respuesta, en las que existe una respuesta objetiva y predefinida. Éstas pueden ser sometidas a un proceso de corrección automática. Por otro lado, preguntas abiertas en las que no hay una respuesta predefinida y existe un cierto grado de subjetividad. En los cursos que poseen este tipo de preguntas, como el que se utiliza de base para este proyecto, se suele utilizar la técnica de revisión entre pares (*peer review*), en la que unos alumnos evalúan el trabajo de otros compañeros. Finalmente, el usuario recibe una nota final en función de los resultados obtenidos en las distintas actividades del curso. Normalmente, realizar el curso no tiene coste alguno pero recibir un certificado sí que puede tener una implicación dineraria. De hecho, desde 2014, las tres plataformas más conocidas (*Coursera*, *edX* y *Udacity*) cobran a los usuarios por la concesión de certificados [6]. La media del coste de éstos por curso está en torno a los 50 \$ (45 € aproximadamente).

El ámbito que concierne a los *MOOCs* es considerablemente nuevo, lo que conlleva que existan relativamente pocos estudios y proyectos de investigación que traten sobre ellos. Respecto a la novedad de estos cursos, se puede considerar a 2012 como “el año de los *MOOCs*” [7], dado que fue en dicho año cuando se fundaron las plataformas más famosas mencionadas anteriormente: *Coursera*, *edX* y *Udacity*. Para cursos en español es además conocida *MiríadaX*, aunque también posee cursos en portugués e inglés, fundada en el mismo año que las tres anteriores.

Los *MOOCs* disponen de **herramientas colaborativas** (foros de discusión y cuentas en redes sociales, entre otras) en las que los alumnos interactúan entre ellos planteando o resolviendo dudas, compartiendo enlaces de posible interés u otorgando *feedback* a los profesores sobre el material facilitado, entre otras posibles aportaciones. En línea con esto, y de la misma forma que en una clase tradicional puede haber alumnos que sean especialmente activos (participando más que el resto con preguntas o presentándose voluntarios para distintas actividades, por ejemplo), en los foros de discusión de los *MOOCs* puede suceder algo similar. Existe la posibilidad de que haya alumnos que participen más intensamente que el resto en las discusiones, respondiendo a muchas preguntas, comprobando y comentando sobre las respuestas de otros compañeros o informando a los coordinadores (*staff*) sobre algún error en el material disponible, entre

otras posibles acciones como se mencionó anteriormente. De esta forma, los alumnos que realicen más aportaciones serán los mayores contribuidores (*top contributors*) del curso. Como se verá más adelante, puede resultar interesante identificar de manera temprana a los estudiantes que formarán este grupo. Adelantando ya pinceladas sobre estos motivos, los *top contributors*, al desarrollar a menudo acciones propias de los profesores, pueden ser de ayuda para mejorar el apoyo dado a los participantes en *MOOCs*. También pueden ayudar a mantener un ambiente agradable y correcto en los foros de discusión [1].

III.2 Objetivos

Este Trabajo de Fin de Grado estudia un *MOOC* organizado por la Universidad Carlos III de Madrid que tuvo lugar entre abril y junio del año 2015, utilizando la plataforma *edX*. Dicho *MOOC* tenía como título ***Introduction to Programming with Java Part 1: Starting to Program with Java*** (traducido, *Introducción a la Programación con Java Parte 1: Empezando a Programar con Java*). A partir de los datos facilitados por dicha plataforma sobre las notas finales de los alumnos y sus actuaciones en el foro de discusión, se pueden plantear dos objetivos.

Por una parte, para este caso de estudio concreto, se puede comprobar si se cumplen o no las **hipótesis planteadas en otros artículos científicos** que estudian otros *MOOCs*. De estas hipótesis y demás conclusiones se habla más adelante en el Capítulo 2. Haciendo brevemente algún adelanto, una hipótesis que destacar podría ser la posible correlación entre los usuarios más activos y los que mejores notas obtienen, y una conclusión relevante podría ser la de no tener en cuenta solo el número de aportaciones de un usuario sino también la calidad de éstas. La utilidad de los resultados que se obtengan podrá facilitar la creación de modelos generales para ayudar a comprender estos cursos y contribuir a mejorar el desarrollo de futuros *MOOCs*. Como ejemplos se podrían mencionar el entender las relaciones entre los estudiantes a través del foro de discusión, predecir el porcentaje de usuarios que finalizarán los cursos, comprender las motivaciones que llevan a los estudiantes a continuar o abandonar un curso o descubrir si los mayores contribuidores responden a cualquier alumno o solo a un cierto grupo.

Por otro lado, de lo mencionado con anterioridad se puede deducir que los denominados ***top contributors*** suponen un papel muy importante en los cursos, ya que en algunos casos desempeñan tareas propias de los coordinadores. De esta forma, resultaría interesante poder **identificarlos de manera temprana** e incluso otorgarles algún privilegio de moderador. Efectivamente, esto se trata de un problema de predicción que sería interesante resolver lo más pronto posible tras comenzar el curso. Como está reflejado en el Capítulo 2, hay varios impedimentos que dificultan esta predicción, como la gran cantidad de mensajes presentes al inicio del curso.

III.3 Marco regulador

Durante la realización de este trabajo se tuvo acceso a tres archivos relacionados con el *MOOC*: uno de ellos contenía los *tweets* que poseían el *hashtag* del curso

(#javaedxuc3m), otro poseía los mensajes del foro del mismo y el otro tenía las notas obtenidas por los alumnos. Los tres archivos tenían varios campos además del texto del propio mensaje, como el nombre de usuario del autor (en el Capítulo 3 se verá una tabla con los campos relativos a los mensajes del foro de discusión). En el tercer archivo también figuraban los nombres reales de los usuarios, por lo que es necesario hablar sobre la normativa relacionada con el manejo de datos personales; esto es, hablar sobre la **Ley Orgánica de Protección de Datos** [8].

La Ley que se acaba de mencionar tiene por finalidad proteger y garantizar los derechos de las personas de las cuales se poseen datos personales. El control establecido por esta Ley permite evitar que, mediante el tratamiento de sus datos, se llegue a disponer de información sobre las personas que afecte a su intimidad y libertades públicas. Alguna declaración que se puede resaltar es, por ejemplo, la que dice que la utilización de información de carácter personal debe estar acorde con la finalidad para la que esta información fue recogida en su momento. También se puede destacar que las personas que ceden sus datos personales han debido ser previamente informadas y cerciorarse de que dan su consentimiento. Ello se relacionará después con la política de privacidad y los términos de uso de la propia plataforma *edX*.

Atendiendo a la **Guía de Seguridad de Datos** [9] se puede establecer el nivel que clasifica a los datos personales de los que se disponen en este proyecto. Según este nivel, se podrán exigir unas medidas de seguridad u otras al tratamiento de los datos. Existen tres niveles: alto, medio y básico. Atendiendo a la explicación que se da en la mencionada guía sobre cada nivel, se concluye que los documentos de los que se disponen en este proyecto pertenecen al nivel básico.

Como se anunció anteriormente, además de a la Ley de Protección de Datos, se debe atender a lo que se dice en la propia página web de *edX*. Por un lado, respecto a los **términos de uso** [10], se puede referenciar la sección “publicaciones de usuario” (*user postings*). En ella se estipula que, al escribir un comentario, el usuario autoriza a *edX* de forma irrevocable a publicar, transferir y distribuir (entre otras acciones) estas publicaciones. Igualmente, también se permite a cada usuario de la web *edX* acceder a los mensajes para sus propios propósitos. Por otro lado, en cuanto a la **política de privacidad** [11], en la sección “nombres de usuario y publicaciones” (*usernames and postings*) se dice que los comentarios y otra información publicada por los usuarios pueden ser accedidos por los que visiten el sitio *edX*, por lo que se pide cuidado a la hora de aportar información que pueda identificar a la persona que hay detrás del nombre de usuario. Además, a la hora de compartir información con terceros, se dice que se podrá destinar, entre otras razones, para investigaciones científicas, sobre todo las que tengan que ver con la educación. Ésta es la que aplica en nuestro caso puesto que este proyecto tiene una componente de investigación sobre *MOOCs* y educación a partir de un curso concreto de esta plataforma. De nuevo, aunque también se tuvo acceso a los nombres reales de los usuarios, no se utilizó más información personal que sus seudónimos, identificadores, mensajes y nota obtenida.

Por todo ello, se justifica la recopilación y el uso de los datos utilizados para el desarrollo de este trabajo respetando el marco regulador existente.

III.4 Metodología

Primero, el que suscribe este proyecto, tras estudiar el listado y la descripción de los trabajos de fin de grado apuntados en la página web de la universidad, tuvo una reunión con los coordinadores este proyecto. De esta manera, se comenzó por leer varios artículos proporcionados por dichos coordinadores. Asimismo, el alumno buscó y leyó algunos artículos adicionales. Los artículos proporcionados por los profesores trataban como caso de estudio otro *MOOC* (llamado “Educación Digital del Futuro”) y apuntaban conclusiones de su investigación. El propósito de la lectura de los mismos fue apuntar de qué datos se partía y qué líneas se podían investigar. Cabe destacar que el alumno ya poseía algunos conocimientos sobre *MOOCs* puesto que en su tercer año de carrera cursó una asignatura en la que se trató sobre ellos (concretamente, la asignatura se llamaba “Conocimiento Libre y Aprendizaje en la Web”).

Tras esto, y una vez el alumno decidió escoger este proyecto, los coordinadores le facilitaron el documento de los *tweets* ya mencionado anteriormente. A partir de este archivo se creó una tabla que indicaba el recuento de dichos *tweets*, asignando a cada usuario el número de mensajes escritos por él. A continuación, tras una nueva reunión entre los coordinadores y el alumno para comentar de nuevo el proyecto y hablar sobre la clasificación anterior, los primeros facilitaron al segundo un archivo de texto donde figuraban los mensajes intercambiados en el foro del *MOOC* que se estaba estudiando. A partir de ese momento, el método de trabajo que se siguió fue de reuniones programadas para que el encargado mostrara los avances que había ido consiguiendo. En dichas reuniones se proponían nuevas características que investigar y se discutían los diferentes acercamientos que se podían realizar a los resultados obtenidos. Esto es, el proceso puede ser clasificado como iterativo (comunicación con los profesores para validar los resultados obtenidos y proponer nuevas ideas) e incremental (se comienza por análisis sencillos sobre los datos disponibles para ir avanzando hacia gráficas o algoritmos más complejos).

Finalmente, durante la realización de la memoria, mediante correo electrónico el alumno fue compartiendo sus avances con los coordinadores y éstos iban sugiriendo mejoras para la misma. Anteriormente, se había discutido cuál podría ser una adecuada estructura de esta memoria, la cual será desarrollada en la siguiente sección. Cabe destacar que la escritura de dicha memoria se ha ido realizando de forma continua e iterativa desde el inicio del proyecto.

III.5 Estructura de la memoria

La memoria comienza con un capítulo previo. En éste se presenta una introducción al trabajo y se menciona brevemente el contexto relacionado con el mismo. Una vez cubierto este tema, se pasa a los objetivos principales que persigue este proyecto, seguidos de una referencia al marco regulatorio relativo a los datos facilitados por la plataforma *edX* y de un breve comentario sobre la metodología seguida en el desarrollo de este proyecto.

Tras el capítulo anterior, en el segundo capítulo se trata el contexto relacionado con una mayor profundidad que en la introducción, al mismo tiempo que se apuntan datos relevantes para este proyecto extraídos de otros documentos. También quedan reflejadas las tecnologías utilizadas durante la realización del trabajo; algunas eran ya conocidas (como *MATLAB*) mientras que ha sido necesario aprender sobre otras (como *Gephi*).

Siguiente, en el tercer capítulo se concreta más en el caso de estudio sobre el que trata este trabajo, siendo éste un *MOOC* llevado a cabo por la Universidad Carlos III entre abril y junio de 2015. De este modo, se explica a qué retos y problemas se enfrenta el proyecto. También se comenta cuál era el diseño y las características del curso y qué estructura tenían los mensajes del foro de discusión en el archivo proporcionado por *edX*.

En el cuarto capítulo, se explican los procedimientos llevados a cabo para solucionar los problemas del capítulo anterior y las acciones realizadas para alcanzar los objetivos propuestos en el capítulo introductorio. Se presenta una tabla (Tabla 4) con los nombres de los algoritmos de *MATLAB* creados junto con una breve descripción de la funcionalidad de cada uno. Adicionalmente, tras la tabla, están desarrollados todos estos algoritmos mediante la utilización de pseudocódigo.

A continuación, en el quinto capítulo se muestran los resultados obtenidos y se discuten brevemente. Primero se aportan resultados enfocados a todo el curso en general y después se muestran resultados más centrados en torno a los *top contributors*.

Como último capítulo tenemos las conclusiones que se pueden sacar a partir de los resultados del capítulo anterior, así como aportar ideas sobre líneas de trabajo futuras. Además, también se dedica un apartado a contemplar las habilidades y competencias que el alumno ha obtenido durante la realización de este proyecto.

Como apéndices se encuentran una planificación del desarrollo de este trabajo y un presupuesto estimado. También aparece una versión en español del capítulo introductorio y del capítulo de conclusiones. Finalmente, aparece un resumen en inglés de todo el proyecto.

Anexo IV

Conclusiones y líneas futuras

IV.1 Conclusiones

Antes de comenzar con las conclusiones, se recuerda los **objetivos principales**: por un lado, comprobar si las hipótesis planteadas por otros artículos y las conclusiones a las que éstos llegaron se cumplen para este caso de estudio; por otro lado, estudiar las características de los mayores contribuidores e intentar predecir de forma temprana qué usuarios llegarán a convertirse en *top contributors*. De la misma manera que los resultados fueron organizados primero con características generales al curso y después con características de los contribuidores más activos, se hará la misma clasificación en este capítulo. También se recuerda que las conclusiones mostradas a continuación son fruto del estudio de un *MOOC* concreto titulado *Introduction to Programming with Java Part 1: Starting to Program with Java*.

Primero se mostrarán brevemente los aspectos más relevantes sobre el uso de la red social *Twitter*. Es bastante claro que el foro de discusión del curso ha sido mucho más utilizado y relevante que los *tweets*. En uno de los trabajos relacionados [13] ya se destacaba que el uso de *Twitter* fue menor que el del foro en el *MOOC* “Educación Digital del Futuro”. En cuanto a la temática de los mensajes (Tabla 3), el 47,7% de ellos trataban sobre comunicación o difusión del curso. Algo similar ocurrió en el *MOOC* que se acaba de citar, aunque en ese caso el porcentaje fue menor debido a que hubo varios

ANEXO IV: CONCLUSIONES Y LÍNEAS FUTURAS

mensajes que no tenían que ver con el curso, lo cual no ha ocurrido en este caso. Por todo ello se concluye que *Twitter* es más útil como herramienta para dar a conocer el curso que como fuente de relaciones entre los usuarios; en cualquier caso, sería conveniente seguir comprobando estos resultados con los de otros *MOOCs*.

Como rasgo **general** del curso, solamente aprobaron el *MOOC* el 2% (1522) de los alumnos matriculados (Figura 3). Esto cumple la premisa del Capítulo 2 que decía que el número de alumnos que superan un *MOOC* es menor al 10% de los que se apuntan a él. Realizando una comparación, en el curso “Educación Digital del Futuro” [13] aprobaron un 8% de los matriculados (456 alumnos). Aunque los porcentajes y valores son distintos debido a que la cantidad de matriculaciones en ambos cursos difiere bastante, se ve que en ambos casos aprueban menos del 10% de los que se apuntan, como se acaba de decir.

En este curso, el 6% del total de alumnos comentó en el foro y el 94% restante no (Figura 3). Se calculó que suspendieron el 95% de los alumnos que no comentaron en el foro (Figura 5). Esto va en consonancia con el hecho de que la gran mayoría de los alumnos registrados no participa en el curso (ni en las discusiones ni en las actividades). En cuanto al 5% restante de los suspensos, la mayoría de éstos serán estudiantes que saludaron en la discusión de bienvenida pero después se despreocuparon del curso, lo cual parece ser corroborado por la Figura 15. Por otra parte, en cuanto a los aprobados, resulta muy interesante que la mitad realizara aportaciones en el foro y que la otra mitad no (Figura 4). Esto es notablemente llamativo ya que cabría esperar una alta correlación positiva entre aprobar y comentar; es decir, que hubiera una cantidad mayor de alumnos que aprobaron y comentaron que de alumnos que no comentaron y aprobaron. Aún así, si se observan todos los datos a la vez (Figura 3), según la información obtenida en este caso de estudio, se pueden realizar las siguientes predicciones: si un alumno ha comentado en el foro, hay un 17% de probabilidad de que apruebe y un 83% de que suspenda; si un alumno no ha realizado aportaciones al foro, existe una probabilidad del 1% de que apruebe y una probabilidad de 99% de que suspenda. Expresado de forma matemática con probabilidades condicionadas:

$$P(\text{aprobar}|\text{comentar}) = \frac{1}{2} = 50\% \qquad P(\text{comentar}|\text{aprobar}) = \frac{1}{6} \approx 17\%$$

En cuanto a las notas obtenidas por los usuarios (Figura 6), como ya se mencionó en el Capítulo 5, se deduce que una gran cantidad de los participantes que no sacan un cero realizan solamente una actividad de evaluación, probablemente la primera, antes de dejar de seguir el curso. También se vio anteriormente que la correlación obtenida entre la cantidad de comentarios publicados y la nota final obtenida es de 0,21. Aunque positiva, es una cantidad baja, por lo que no se pueden extraer conclusiones de ella. Lo mismo ocurre con la relación utilizada entre hilos iniciados y comentarios en discusiones ya creadas, cuya correlación con las notas es de 0,25.

Siguiendo con las conclusiones, la gráfica que representa la evolución del número de comentarios conforme al transcurso del *MOOC* (Figura 7) se corresponde con lo que ocurre normalmente en estos cursos (de nuevo, por la emoción que se produce al inicio del mismo y por las discusiones de presentación y bienvenida). Esto es, una gran cantidad de mensajes los primeros días que siguen posteriormente con un pronunciado descenso, y después alcanzan picos de aportaciones sobre todo en fechas de liberación de material. En concordancia con este aspecto y con la clasificación anterior de aprobaros y suspensos, se

tiene que la gran mayoría de los estudiantes (3699, un 68% del total) realizaron solamente una aportación (Figura 8). El número de participantes que realizaron dos comentarios se reduce a 689 (13% del total). La media es de 2,5 mensajes por usuario.

Respecto a las franjas horarias en las que hay mayor cantidad de mensajes, obviamente ha de tenerse en cuenta la diferencia horaria entre distintas partes del mundo (Figura 11). Es este estudio se ha visto que las tardes de Estados Unidos y de Europa tienen un peso importante mientras que la tarde de India no tiene tanta relevancia (se utilizan estos lugares como referencias ya que gran parte de los alumnos pertenecían a ellos).

Se ha visto anteriormente que los datos del caso de estudio sobre el que trata este proyecto señalan que el 6% de los participantes matriculados realizaron alguna aportación en los foros, aunque solamente el 1% del total comentó y aprobó. Siguiendo con este aspecto, se ha calculado que, cuantas más aportaciones haya realizado un alumno, más probabilidad se tiene de que éste haya aprobado el curso (Figura 13). Intuitivamente, esto se corresponde con la implicación de los estudiantes en el *MOOC*. Por ejemplo, en los participantes que realizaron al menos 8 aportaciones la incertidumbre de aprobar o suspender está al 50%. A partir de los 3 comentarios, la gráfica se aproxima a una recta (hasta alcanzar el 100% de alumnos aprobados). Algo similar ocurre si solamente se tienen en cuenta las dos primeras semanas del curso, aunque la aproximación a una recta es menos fiable. En cuanto a los datos de todo el curso, se puede obtener de manera sencilla la ecuación de la recta que se ha mencionado (recordando que no se debe aplicar a un porcentaje menor al 30%):

$$\left. \begin{array}{l} P(\text{aprobar} | \text{comentar} > 9 \text{ veces}) = 50\% \\ P(\text{aprobar} | \text{comentar} > 24 \text{ veces}) = 75\% \\ P(\text{aprobar} | \text{comentar} > 36 \text{ veces}) = 100\% \end{array} \right\} \rightarrow \% \text{ Aprobados} \approx 1,9 \cdot \text{NumComentarios} + 32,9$$

Respecto a los alumnos que aprobaron el curso y comentaron en el foro, se hallaron datos sobre cuándo realizaron su última aportación a las discusiones (Figura 16). Para cuando expiró el intervalo disponible para realizar la última actividad de evaluación, aproximadamente un 20% de los que aprobaron y comentaron seguían realizando aportaciones (Figura 17). Cabe notar que, aunque el porcentaje de alumnos que continúan comentando va decreciendo conforme avanza el curso, hay una caída más pronunciada en torno al 2 de junio, última fecha en la que se liberó nuevo material. Así, es posible que varios alumnos realizaran la evaluación de dicho material poco después de que estuviera disponible y ya no entraran más al foro dado que el *MOOC* habría terminado para ellos.

Antes de aportar datos concretos sobre los mayores contribuidores, se recuerda que la temprana identificación de éstos está relacionada con otorgarles permisos especiales de moderadores en los foros (la plataforma *edX* los llama *Community Teaching Assistant* o *Community T.A.*). Concretando los datos obtenidos en los **top contributors**, 17 de ellos realizaron aportaciones al menos en un 25% de los días que duró el curso. Algunos de ellos realizaron comentarios de manera más continuada y otros de manera más esporádica. Además, la identificación de estos contribuidores de manera temprana resulta complicada según los datos numéricos obtenidos. Por ejemplo, 5 de los 20 *top contributors* no realizaron ninguna aportación durante las dos primeras semanas del curso (tal vez fueran *latecomers*, lo cual resaltaría la importancia de este grupo como en otros

estudios [13]), y otros 6 contribuidores hicieron menos de 5 comentarios (Tabla 6). Si se realiza una ordenación de mayor a menor número de comentarios teniendo en cuenta solamente estas dos primeras semanas del curso, y se seleccionan a los 20 usuarios con más aportaciones (eliminando coordinadores y *staff*), entre ellos se encuentran 8 de los “verdaderos” *top contributors* de todo el curso (concretamente, los *top* 1, 2, 3, 6, 8, 12, 13 y 20). Así, si de manera temprana se seleccionara a los 10 primeros, se habrían cogido a 5 de los “verdaderos” *top contributors*.

En cuanto a sus mensajes en los foros, en este caso, los 10 mayores contribuidores han respondido a más hilos de los que han creado (Tabla 7); solamente en 3 ocasiones el porcentaje de respuestas es menor del 75% y el de inicios es mayor del 25%. Hay un caso en el cual el porcentaje de inicios de discusión es del 44,6% del total de mensajes publicados; este usuario creaba hilos de discusión en los que informaba a los coordinadores de algún error o en los que pedía información sobre la fecha de publicación de algún material.

En cuanto a los enlaces que forman los contribuidores más activos con otros usuarios (Figura 23), se puede concluir que estos *top contributors* no parecen centrarse en un pequeño grupo de usuarios. Así, aunque hay algunos participantes que solamente están unidos a un *top contributor*, muchos de ellos poseen uniones con varios. Incluso hay uniones entre los propios mayores contribuidores.

IV.2 Competencias adquiridas

Como encargado del proyecto, he adquirido habilidades y competencias durante la realización de este trabajo, como ocurre con cualquier otra asignatura.

Tal vez el concepto más notable a destacar sea la diferencia que el Trabajo de Fin de Grado guarda con respecto al resto de asignaturas, ya que, si no me equivoco, éste ha sido el primer caso en el que me he enfrentado a un problema sin un profesor que semanalmente me fuera enseñando cosas nuevas (aunque han existido varias reuniones presenciales con los coordinadores). En esta línea, por primera vez el problema en cuestión no tenía una solución que pudiésemos comprobar en un libro de soluciones, por ejemplo. Efectivamente, se ha tratado de un proyecto con una componente de investigación porque no sabíamos lo que obtendríamos. En cierto modo, algo similar ocurre en otras asignaturas cuando se nos encarga algún tipo de proyecto o trabajo, pero igualmente la diferencia reside en que, en esos casos, con el transcurso del tiempo (y de la asignatura) obtenemos más conocimientos con los que ir completando dicho proyecto. Además, he tenido que leer varios artículos científicos como trabajo inicial, muchos de ellos en inglés. Este aspecto, además de servirme como punto de partida y ver qué investigaciones se han realizado, me ha aportado más cultura general sobre los *MOOCs* y ha ampliado mis conocimientos sobre el funcionamiento y las características de los mismos.

Como complemento a lo que se acaba de citar, de nuevo ha sido ésta la primera vez que me he enfrentado a un problema de organización y clasificación de información. Para ello, además de aprender algunas funcionalidades que no conocía sobre *Microsoft Excel*, he utilizado algoritmos que he programado con *MATLAB*, una herramienta que aprendí a

usar en el primer curso de la carrera y que hemos seguido utilizando a lo largo de la misma. De este modo, realmente considero que he puesto a prueba mis conocimientos en programación. Como pequeña curiosidad anotaré que todo el tema de programación me apasionó especialmente desde que tuve la primera asignatura el primer año de carrera, por lo que me he enfrentado a este reto con mucho gusto e iniciativa.

Con todo ello, y finalizando esta sección, considero que he obtenido unos resultados satisfactorios en cuanto a contribuir a mi formación como estudiante y como persona.

IV.3 Líneas futuras

Como trabajo futuro, principalmente se puede seguir con el **estudio de los *top contributors*** en varias líneas. Dentro de este mismo *MOOC*, se pueden hacer algunas comprobaciones adicionales, como examinar el grado de tecnicidad de las aportaciones de los usuarios más activos. En la misma línea, se podría comprobar si cada uno se centra en comentar siempre sobre un mismo tema o no. Otro aspecto que no se ha tenido en cuenta durante la realización de este trabajo es el perfil de los participantes. Así, si se observara más información personal de los mayores contribuidores, tal vez existiera un rasgo común o patrón entre ellos (sobre todo puede ser relevante la formación previa de los participantes).

En líneas más técnicas y aspectos de sencilla automatización, nuevos algoritmos pueden ser programados. Quizá pueda ser interesante la distribución horaria de las aportaciones de los usuarios más activos (esto es, obtener una gráfica similar a la Figura 11 teniendo en cuenta solamente a los *top contributors*). También se podrían seleccionar las aportaciones que tienen más votos positivos y comprobar a qué usuario pertenecen. Además, siguiendo con esta línea, el campo *endorsed* de los mensajes podría tenerse en cuenta para un análisis similar (se recuerda que, cuando dicho campo tiene *true* como valor, un moderador o profesor ha calificado a una aportación concreta como relevante). En cuanto a análisis estadísticos algo más complejos, se podría realizar algún estudio de la correlación entre número de aportaciones realizadas y notas obtenidas restringiendo los cálculos a cada semana del curso; esto es, comprobar las notas obtenidas en las actividades de cada semana y relacionarlas con el número de contribuciones relacionadas con esa semana.

Cuando se hayan realizado diversos estudios sobre la identificación de los mayores contribuidores, y la cantidad de información sea lo suficientemente elevada como para poder hacer un grupo de entrenamiento (*train*) y un grupo de prueba (*test*), se podrían realizar estudios mediante regresiones lineales o extracciones de descriptores. En este sentido, se crearía un algoritmo que clasificara de manera adecuada el grupo *train* y después se evaluarían los resultados en el grupo *test*. Si no se obtienen resultados satisfactorios, se modificaría el código y se repetiría el proceso. Tal vez sería interesante tener una probabilidad de detección muy alta a costa de tener una probabilidad de falsa alarma algo elevada (esto es, detectar a prácticamente todos los mayores contribuidores a cambio de seleccionar también a un grupo de alumnos que no llegarían a ser *top contributors*). Además, se podrían ir refinando los resultados haciendo análisis similares cada semana y teniendo en cuenta todos los datos acumulados.

ANEXO IV: CONCLUSIONES Y LÍNEAS FUTURAS

Fuera de este curso, podría examinarse la forma de actuar en otros *MOOCs* por parte de los usuarios calificados como *top contributors* en este proyecto. También podría comprobarse si éstos son habitualmente activos en redes sociales como *Twitter*. Por supuesto, también cabría estudiar si los datos obtenidos en este trabajo se asemejan o no a otros *MOOCs*. Sería interesante dividir éstos entre cursos con temáticas similares al que se ha cubierto en este trabajo y cursos distintos (esto es, puede que las relaciones cambien notablemente entre un curso de programación y uno que trate sobre medicina).

Annex V

Summary

V.1 Introduction and objectives

Massive Open Online Courses (MOOCs) due to their abbreviation in English; *Cursos Masivos en Línea y Abiertos* in Spanish) have captured the interest of many high-level education institutions. Their defenders argue that MOOCs make access to education easier to as many people as possible, increase the visibility of the institution which creates them and allow teachers to experiment with online education to a wide and diverse number of students [1]. On the other side, people who oppose to MOOCs criticize them as virtually harming, as technology disruptive of education that offers a watered-down formal learning and as a risk of even more cuts in government-funded school budgets [2]. Even though this document will not participate in this debate, it should be accepted that MOOCs have considerably impacted on the way of teaching and learning (although they will probably not replace traditional face-to-face teaching, at least for now). Actually, in one of the referenced documents, its authors asked themselves whether MOOCs would replace post-secondary education or not [3]. In that document, MOOCs are considered not as an alternative to higher education but as its complement. Besides, an example is given comparing a computer and a tablet: one is not a substitute of the other; both accomplish different purposes. Thus, the same way MOOCs have features that traditional teaching methods do not, face-to-face education has characteristics which MOOCs lack.

Trying to explain **typical features** of MOOCs [4] [5], they are Internet-offered courses that rely on different platforms like *MiríadaX* or *edX*. A list of offered courses and information about them is typically available to anyone navigating the Web, even though a registration is required in order to enroll on the course. When selecting a MOOC, this will usually show additional information like its duration and start date, among others. In addition, sometimes it can also be found a description of the course and different user ratings if it is not the first time the MOOC is taught, and even a video introducing the course. Once registered on the course and once it starts, different material provided by the organizers will be given, like explicative videos, documents dealing with the concepts and tasks for the student to be evaluated. This student will be able to complete these tasks as he or she wishes as long as it is done before the assignment deadline. Said assignments can be split into two groups. On one side, watching videos and others contents including information about the subject. On the other side, there are exercises, which can be divided in two subgroups. On one hand, close-ended questions like multiple-choice tests in which there exists an objective and predefined answer. These can be passed through an automatic system that corrects them. On the other hand, open questions in which there is not a predefined answer and there exists a certain degree of subjectivity. In MOOCs that contain these types of questions, like the one studied in this document, it is usually utilized a technique called “peer review”, in which a student grades the work done by other users. In the end, each participant receives a final mark according to his or her results throughout the different course activities. Normally, taking the course is free of charge but obtaining a certificate may incur in a monetary expense. Actually, from 2014, the three most famous platforms (*Coursera*, *edX* and *Udacity*) demand payment to users who wish to obtain a certificate [6]. On average, these certificates cost around \$50 (roughly €45).

The field that comprehends MOOCs is considerably new, which implies a lack in studies and research papers addressing them. Regarding the newness of these courses, 2012 was claimed to be “the year of the MOOC” [7] since the most famous platforms aforementioned were founded in that year: *Coursera*, *edX* and *Udacity*. Regarding Spanish courses, the platform *MiríadaX* is also well-know, although it also has Portuguese and English courses. It was founded in 2012, too.

MOOCs have **social tools** (discussion forums and accounts in different social networks, among others) in which the students interact with each others by formulating questions or answering them, sharing possibly interesting links or giving feedback to the teachers regarding the material, amongst other actions. Along these lines, and in the same way in a traditional face-to-face course there may exist students significantly active (by participating more than their classmates or volunteering for different activities, for instance), inside the MOOCs discussion forums something similar may happen. There exists the possibility that some students will become more involved in the forums than their course mates, by checking and commenting about already-posted answers or by reporting errors to the staff more frequently than their peers, among other possibilities mentioned above. In this way, participants that make the most contributions will become the course “top contributors”. As it will be seen later, it may be interesting to identify these contributors early. Bringing forward some of the reasons, as top contributors often carry out actions attributable to teachers, they may result in great help towards other participants. They can also assist in keeping a stable and pleasant environment around the discussion forums [1].

Moving on to the objectives, this Bachelor's thesis studies a MOOC organized by *Universidad Carlos III de Madrid*, which took place between April and June of 2015, using the platform *edX*. The MOOC title was “**Introduction to Programming with Java Part 1: Starting to Program with Java**”. Based on the data provided by said platform regarding student's final marks and their contributions to the forum two objectives can be suggested.

On one side, it can be checked whether the **hypotheses proposed by other scientific papers** addressing other MOOCs are fulfilled or not for this case of study. Some of these hypotheses and other conclusions will be dealt with in Section V.2. Anticipating slightly these issues, a relevant hypothesis could be the one addressing the correlation between the users that contribute the most and the ones that obtain the best marks, and a notable conclusion could be the one that states that it should be taken into account not only the number of contributions but the quality of these, too. The usefulness of these results could allow the creation of models that can be generalized to all MOOCs, in order to help understand these courses and contributing to improve the progress of future ones. As examples, it could be specified understanding how students are connected with each others via the discussion forums, predicting which percentage of students will finalize the courses, understanding the causes by which a participant continues or drops out the MOOC or discovering whether the top contributors answer to any other student or just to a small group of participants.

On the other side, from what was mentioned before it can be deduced that these top contributors play an important role in these courses since they sometimes perform tasks that are expected from teachers and staff. This way, it could be interesting to **identify these top contributors early** in the course and even grant them some kind of moderator benefit. Certainly, this is a prediction problem that would be interesting to solve as soon as possible after the start of the MOOC. As it is specified in Section V.2, there exist some impediments that hinder this prediction, like the large number of messages exchanged in the beginning on the course.

V.2 Context and related work

As another part of this project, at its beginning different recent papers dealing with similar MOOC studies were read. In these papers, some classifications of the users were made and advices to other projects aiming to study MOOCs were provided. Here, the most relevant conclusions from those documents will be collected.

First of all, there is a great **difference** between MOOCs and traditional face-to-face learning, being such difference the large number of registered students (there also exist “closed” online courses that share the number enrolled participants with traditional courses). On the other side, in these traditional courses, the number of matriculated students is generally lesser (a few tens) and it is presupposed that they are willing to face and actively participate in the course, usually due to the high economic cost that the registration requires. However, in MOOCs, since enrollment is usually free, the number of participants ascends to hundred and even thousands (although there may exist economic costs like the one required in order to obtain a certificate). Still, this issue has

an important aspect, which is that the number of participants that pass a course is usually lesser than 10% of the ones that enroll on it [12]. Even though this percentage may seem small, the quantity itself usually concerns hundreds of students, which is generally higher than the number of students that pass in traditional face-to-face courses.

In relation with the previous paragraph and focusing the attention again on MOOCs, the most interested students will usually be the ones contributing more than others to social tools, being the course forum the most used in general. On one hand, this is helpful when trying to identify top contributors (it is reminded that, with this term, the users with the most contributions to the forum are named). On the other hand, the quality of these contributions should be also taken into account since a participant may post many contributions without really adding useful material. Moreover, there is an additional **issue** when trying to identify early these top contributors, being this issue the number of messages posted in the forum during the beginning of the course due to the excitement or hype generated, and even more if a welcoming discussion is opened [13] [1]. Thus, there is a notable noise presence when trying to identify soon these contributors. Another problem for their early identification is the existence of users that enroll on the course once this is already started (called latecomers in some documents [13]). Sometimes [13], an interesting number of users that pass belong to this group (26.1% of students that passed the course “Digital Education of the Future” were latecomers; if this percentage is transformed into specific numbers, 119 latecomers from 456 passing students are obtained [13]). This fact may complicate this kind of studies since there will probably be few contributions from these participants during the first days of the course (in effect, latecomers are characterized by not being present at the beginning of the MOOC) and maybe some of them get to reach top contributions. The aforementioned course, “Digital Education of the Future”, has an important relevance in this work since it is the MOOC used as case of study for two papers used as references [13] [1].

It can also be found some results that are specifically described as **starting point** for projects like this one. Said results were conclusions from studies on other MOOC. Three of them are not to take into account users’ gender, to make predictions once the first evaluation task has taken place and to take into consideration contributions to all social networks instead of making different group for each tool [1]. In addition, it is advisable not to consider only the number of contributions but include characteristics like their quality or in which period they were posted. Another opinion found in some papers [13] [1] is the one that recommends “gamification” aspects. These are about utilizing techniques and methods characteristic of videogames to promote a funny and bearable environment in what *a priori* may seem like a boring activity. These techniques can vary from giving users some kind of badge to allowing them to customize their usernames or profile images with special characteristics. This “gamification” issue influences in participants’ motivation: if they are more motivated, they will probably be more willing to actively get involved in the course, both by completing activities and by contributing to the forum. Effectively, they are small details that may influence the formation of a better and more pleasant experience for the participants.

Coming to an end, some miscellaneous notes can be provided. For example, it could be thought that acquiring a certificate by passing the course does not necessarily mean that the student has learned many things [12] (although this situation should be unlikely unless the user cheats or the course is not well designed). This is, it may happen that a latecomer is not able to get a certificate (since he registered late) but has acquired a large

amount of knowledge, while someone that passes the course and receives a certificate has not learned too much. Another relevant advice is having different social networks in MOOCs for each student to use the one he or she feels more comfortable with, as well as obtaining different participation level (although the course forum is usually the most used tool) [13]. Before finishing, some usefulness of identifying top contributors early in the course will be provided. In platforms like *edX*, teachers can assign certain privileges to these students, like allowing them to edit or delete messages written by their peers [1]. In this way, a collaboration to keep the forum utility is forged and the support given to MOOC participants could be improved by promoting discussion around them.

Briefly, with respect to technologies and used tools, *Notepad++* was used to open the files. Said files had the extension *.mongo*, generated by the database *MongoDB*, which is based in *NoSQL* [23]. To obtain some graphs and arrange data, the software *Microsoft Excel* was utilized. The algorithms were developed and run by using the calculus program *MATLAB*. To generate social graphs, it was tested with both *NodeXL* [27], an open-source plug-in for *Excel*, and *Gephi* [28], software designed for network generation and visualization (finally, the graphs in this document are the ones generated with *Gephi*). To write this work, *Microsoft Word* was used, a program destined to create and edit texts [29]. Finally, the platform *edX* itself may be mentioned along with its content distribution and its forums.

V.3 Case of study

Extending what was mentioned before, this Bachelor's thesis focuses on a MOOC that took place between April 20th and June 30th of 2015 titled “**Introduction to Programming with Java Part 1: Starting to Program with Java**” and published on the platform *edX*.

As general course information which could be read from its page (it cannot be referenced since, once the course finished, the page disappeared; the 2016 version of the same course can be referenced in this moment [31]), it is an introduction to programming in Java language, which is one of most in-demand languages since it is used in a wide number of software platforms. This course significance was remarked stating that it is one of the basic knowledge that should appear in every technology expert's curriculum.

The course in question was the **first part out of three**. This first part was aimed to familiarize students with basic concepts, moving on to more complex ones during the 5 weeks the course lasted. Examples and cases of study were provided for students to implement their own simple programs or collaborating with other users. Part two would focus on how to write “good” code, and the third one would mostly deal with data structures (currently, the second part can be referenced [32]).

Lastly, users were encouraged to enroll on the course by mentioning that knowledge about Programming is needed not only for its use in computers and smartphones, but also for its importance with respect to providing techniques around computational thinking. In the course version of 2016 it can also be seen some user reviews from participants who took the course in 2015. Last time it was checked, this MOOC had a rating of 4.5 starts out of 5 and it had 74 reviews. It is also interesting that, in the critiques section, it can be

checked if the user that wrote the analysis completed the course, dropped it or if he or she is taking it in that moment. The rate given by each individual user can also be seen.

Regarding more technical course parameters, this one had duration of 2 months, approximately (even though the material was divided among 5 weeks) and an effort between 5 and 7 hours per week was estimated. It was a free MOOC that gave the option of obtaining a certificate for \$25 (around €22). It was included in the category Computer Science at an introductory level and it was hosted from *Universidad Carlos III de Madrid* (Telematics and Computing Departments together). The language in which it was taught was English, and the videos had transcripts in English, Spanish, Portuguese and Mandarin Chinese.

With respect to the material, as aforementioned, the course was organized in 5 weeks. Each one of them had explicative videos, texts, graphs, tables and exercises for the student to do. In the version studied in this work (the one from 2015), there were “peer review” activities and the evaluation exams had fixed deadlines (two weeks after being made available, except for the first one, which had three weeks). Effectively, this aspect of fixed deadlines reduces the number of students that pass the course. On the contrary, the version from 2016 did not have such “peer reviews” and the exams were available to take until the MOOC ended. Regarding the scoring system, there were 5 **exams** of 15 points each (one per week) and 2 **peer review** activities of 10 and 15 points in weeks 3 and 5 respectively. Thus, the total score would vary between 0 and 100 points, and a minimum mark of 60 points was needed in order to pass the course. Respect to implications derived from the “peer review” activities, some can be mentioned [33]. For instance, it can be found a lack of teacher supervision due to the large number of activities to evaluate (this large number of activities also affects the coordination to link tasks and evaluators). Another implication is these tasks degree of subjectivity and the little to no experience some users may have with correcting exercises. This aspect may expound itself in the form of complaints about the received marks (some *Twitter* and discussion forum messages were about this issue). It should be pointed out that, in courses that have “peer review” assignments, the number of participants that complete the course is lesser [33].

V.4 Solution design and implementation

During this work, the tool *MATLAB* was used to program and execute several algorithms. This files extension is *.m* and they are programmed in a language that is similar to *Java* or *C*. In Table 10 it is shown the name of each script and a description of what each one accomplishes.

Name	Average execution time	Number of lines of code	Description
<i>Certificados</i> (Algorithm 1)	20 s	123	Calculates the number and percentage of students passing and failing the course according to whether they contributed to the forum or not (Figure 27)

ANNEX V: SUMMARY

<i>Certificados_2semanas</i> (Algorithm 1)	15 s	116	Similar to <i>Certificados</i> but restricting the calculations to the first two weeks of the course (Figure 28)
<i>ClasificacionDiagramaSocial</i> (Algorithm 2)	174 s (2.9 min)	49	Classifies the messages in pairs “replier-replied”
<i>Histogramas</i> (Algorithm 3)	834 s (13.9 min)	115	Both for all the course and the first two weeks, represents a histogram with “number of messages respect to number of users with X number of messages” (Figure 29) The histograms are also divided in intervals of different scale (Figure 30 and Figure 31)
<i>RecuentoCodigo</i> (Algorithm 4)	Op. 1: 576 s (9.6 min) Op. 2: 576 s (9.6 min) Op. 3: 14 s Op. 4: 6 s	149	Allows to choose among 4 options Option 1: counts the number of comments published by each user, identifying these with their “author_username” Option 2: counts the number of comments published by each user, identifying these with their “author_id” Option 3: counts the number of published messages each day (Figure 44) Option 4: counts the number of published messages in each hour of the day for all the course and eliminating the first three days (Figure 32 and Figure 33, respectively)
<i>RecuentoCodigo2semanas</i> (Algorithm 4)	Op. 1: 255 s (4.25 min) Op. 2: 258 s (4.3 min)	27	Similar to <i>RecuentoCodigo</i> but restricting the calculations to the first two weeks of the course and without options 3 and 4
<i>RestriccionDiagramaSocial</i> (Algorithm 5)	4 s	25	From the results obtained in <i>ClasificacionDiagramaSocial</i> and the number of comments of each user, makes a “replier-replied” classification selecting only users that replied more than an specified number of times (Figure 34 and Figure 35)

<i>UltimoComentarioAprobados</i> (Algorithm 6)	17 s	138	From the students that passed the course, selects the date of their last registered contribution (the same can be done with the participants that failed by varying a parameter; Figure 36, Figure 37 and Figure 38)
<i>PlotsTops</i> (Algorithm 7)	7 s	71	Regarding the 20 top contributors, counts the number of messages that each one posted each day (Figure 39, Figure 40, Figure 41 and Figure 42)
<i>Correlaciones</i> (Algorithm 8)	3.9 s	48	Calculates the correlation between users' final marks (excluding the ones that obtained a 0) and the number of contributions they made. Also calculates the correlation between users' final mark and a established relation between started discussions and comments in already created threads
<i>Notas</i> (Algorithm 9)	12 s	32	For the whole range of final marks (excluding 0), calculates the number of students that obtained each mark (Figure 43)
Name	Average execution time	Number of lines of code	Description

Table 10. Scripts designed with *MATLAB*.

ANNEX V: SUMMARY

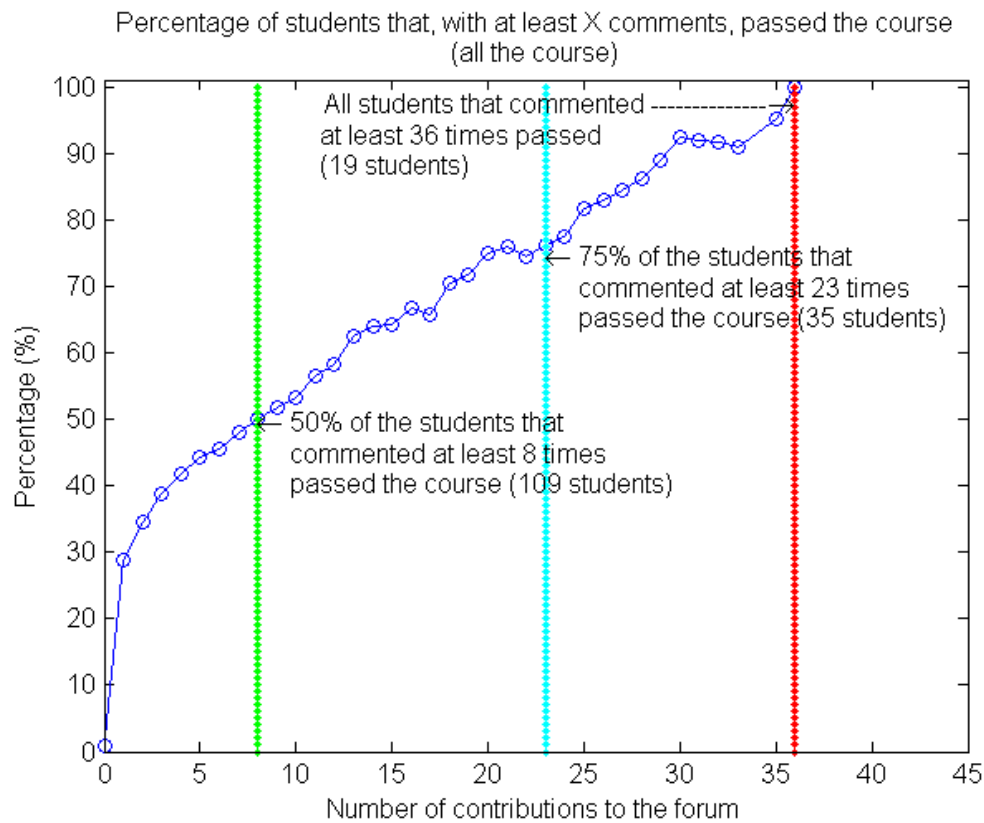


Figure 27. For the whole course, percentage of students that, with at least a given number of contributions to the forum, passed the MOOC.

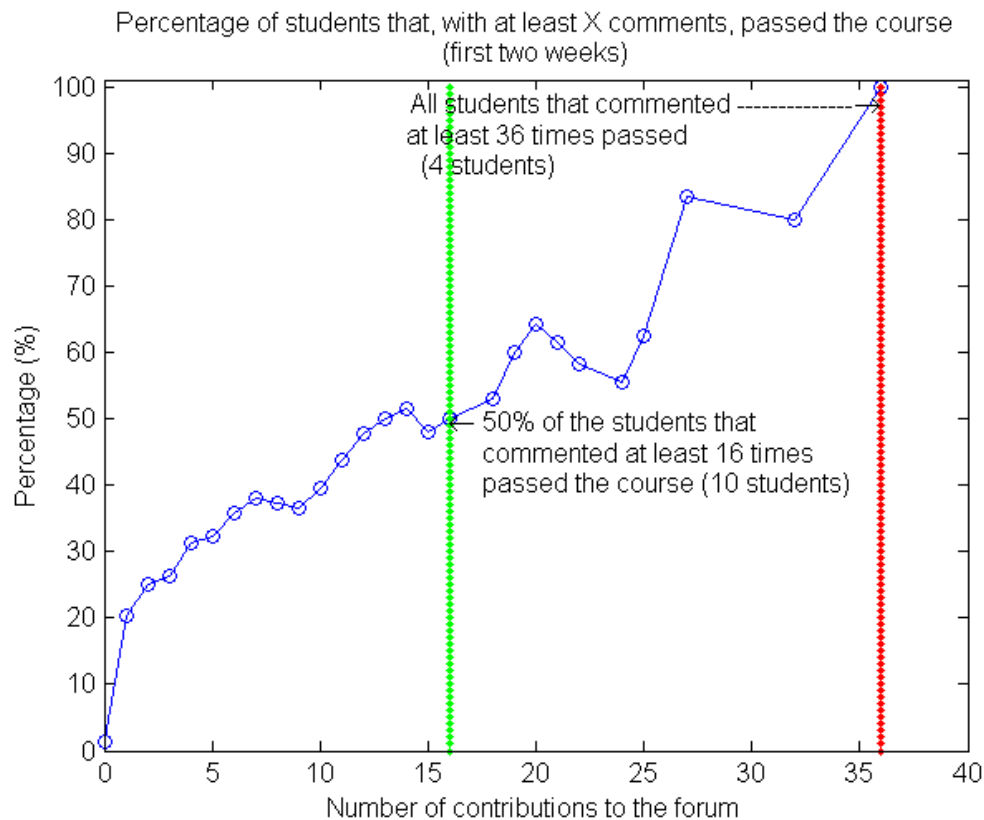


Figure 28. For the first two weeks of the course, percentage of students that, with at least a given number of contributions to the forum, passed the course.

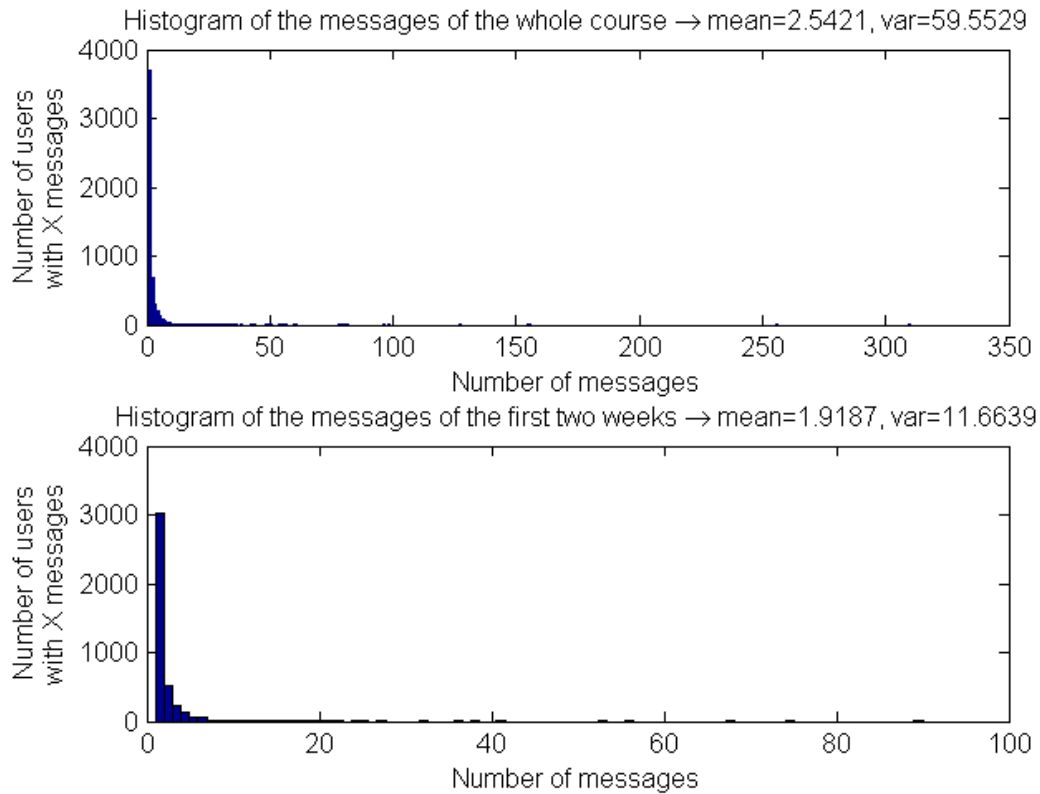


Figure 29. Histograms representing the number of users that contributed with different number of messages to the forum.

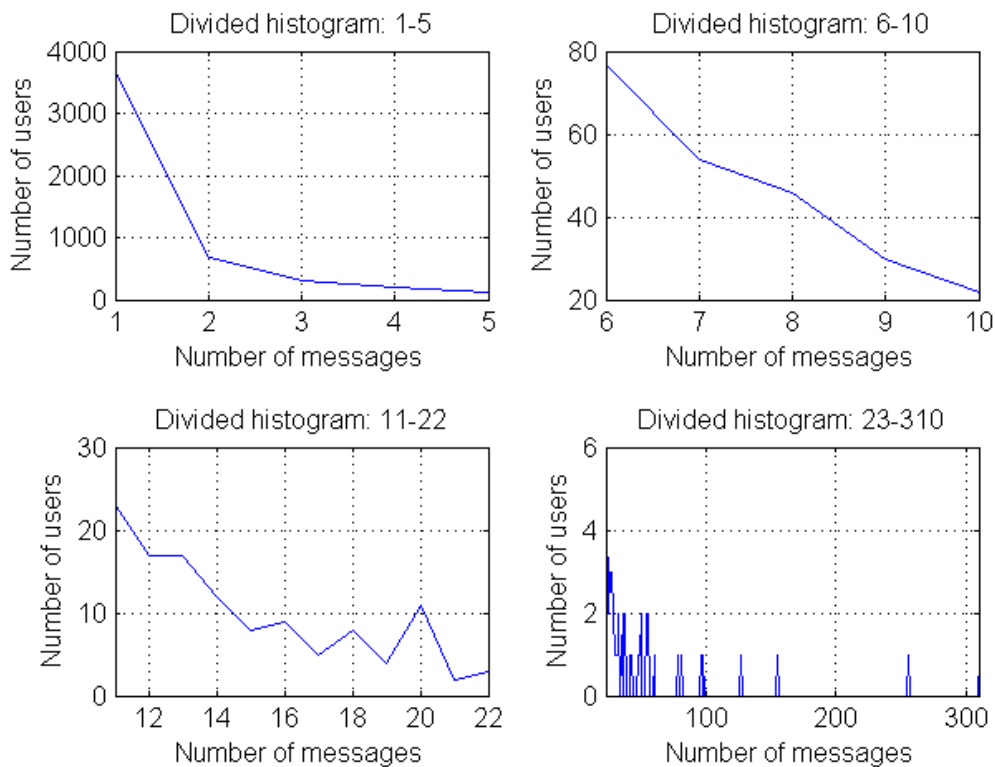


Figure 30. Histogram divided in intervals with different lengths for a better interpretation (the whole course). Note: moderators and staff were not excluded.

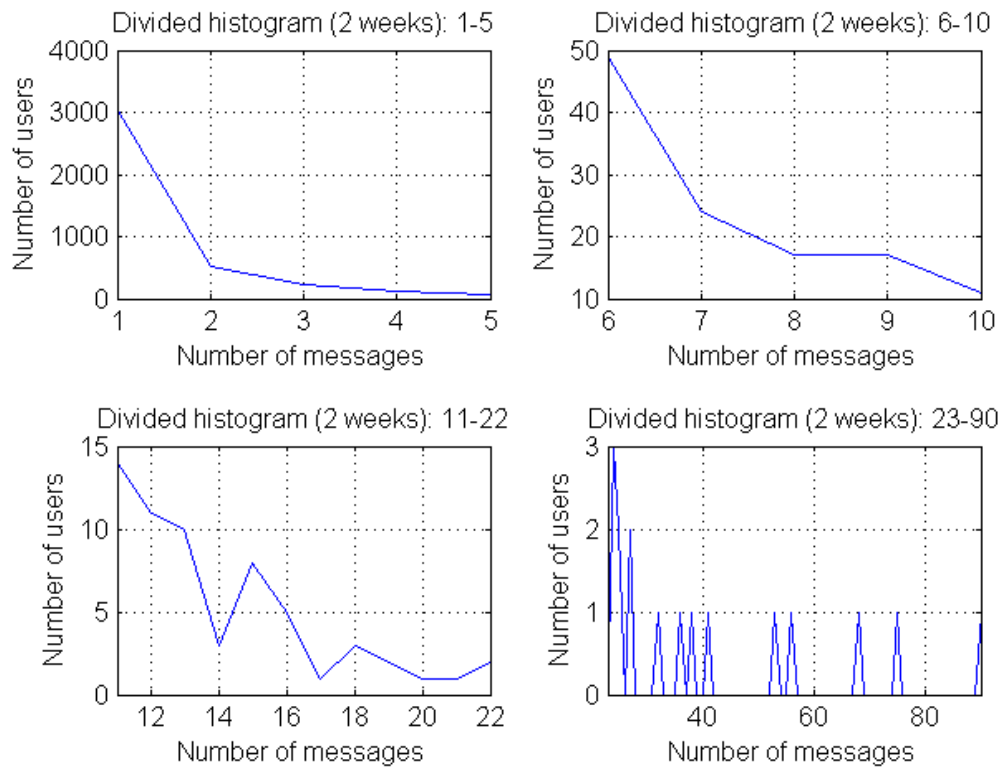


Figure 31. Histogram divided in intervals with different lengths for a better interpretation (first two weeks of the course). Note: moderators and staff were not excluded.

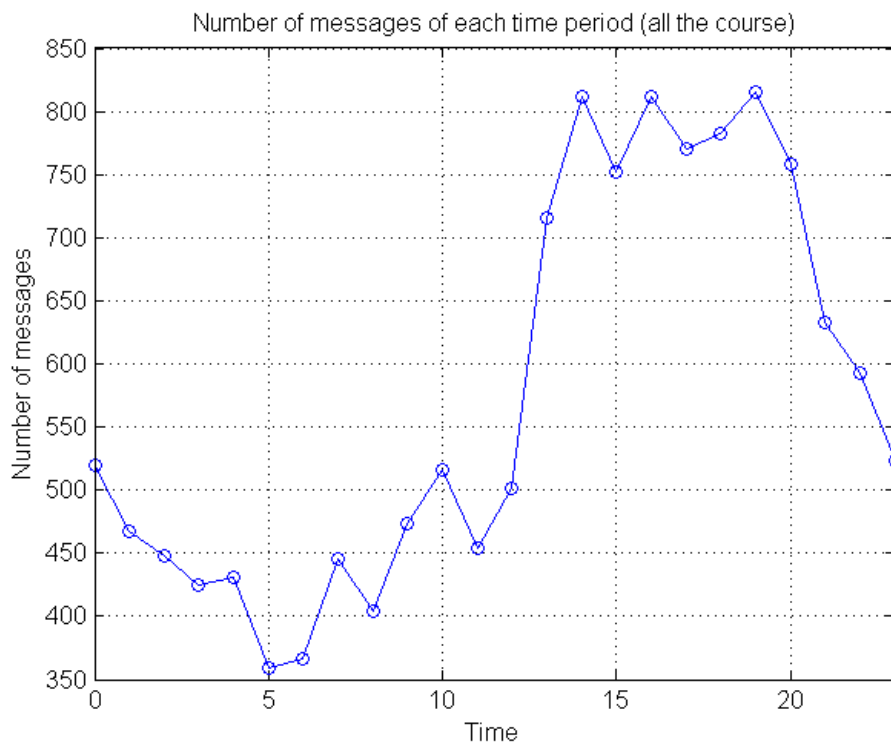


Figure 32. Number of messages in each time period (the whole course; UTC format). Note: the value corresponding to a time, X, is the number of contributions between X:00:00 and X:59:59.

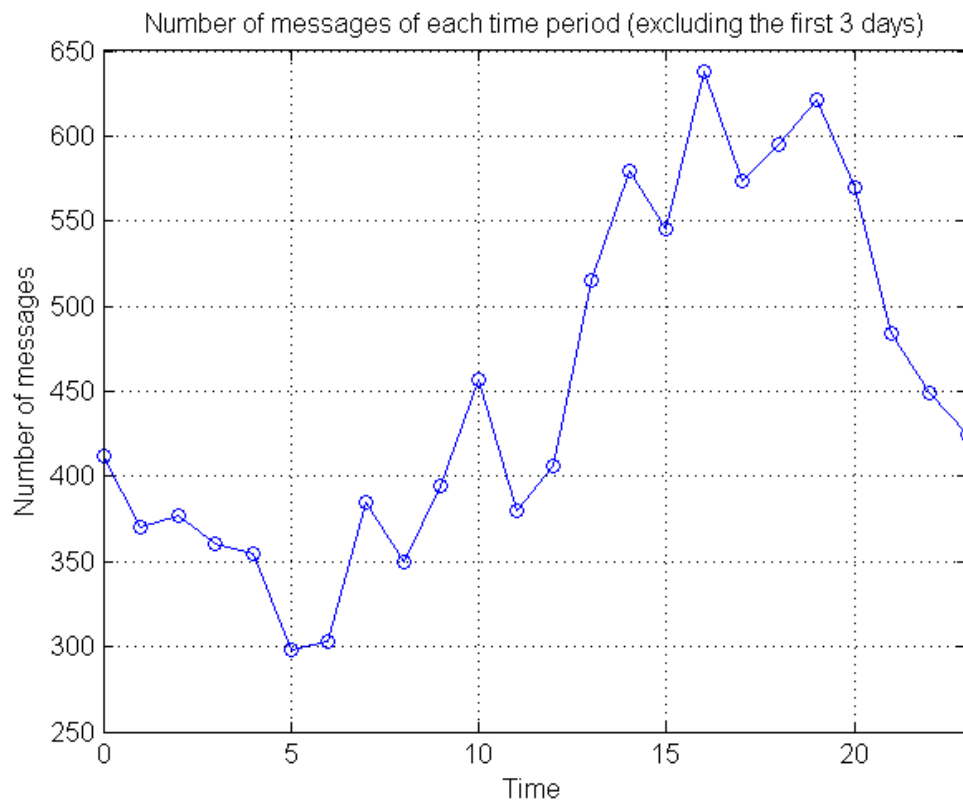


Figure 33. Number of messages in each time period (excluding the first 3 days; UTC format). Note: the value corresponding to a time, X, is the number of contributions between X:00:00 and X:59:59.

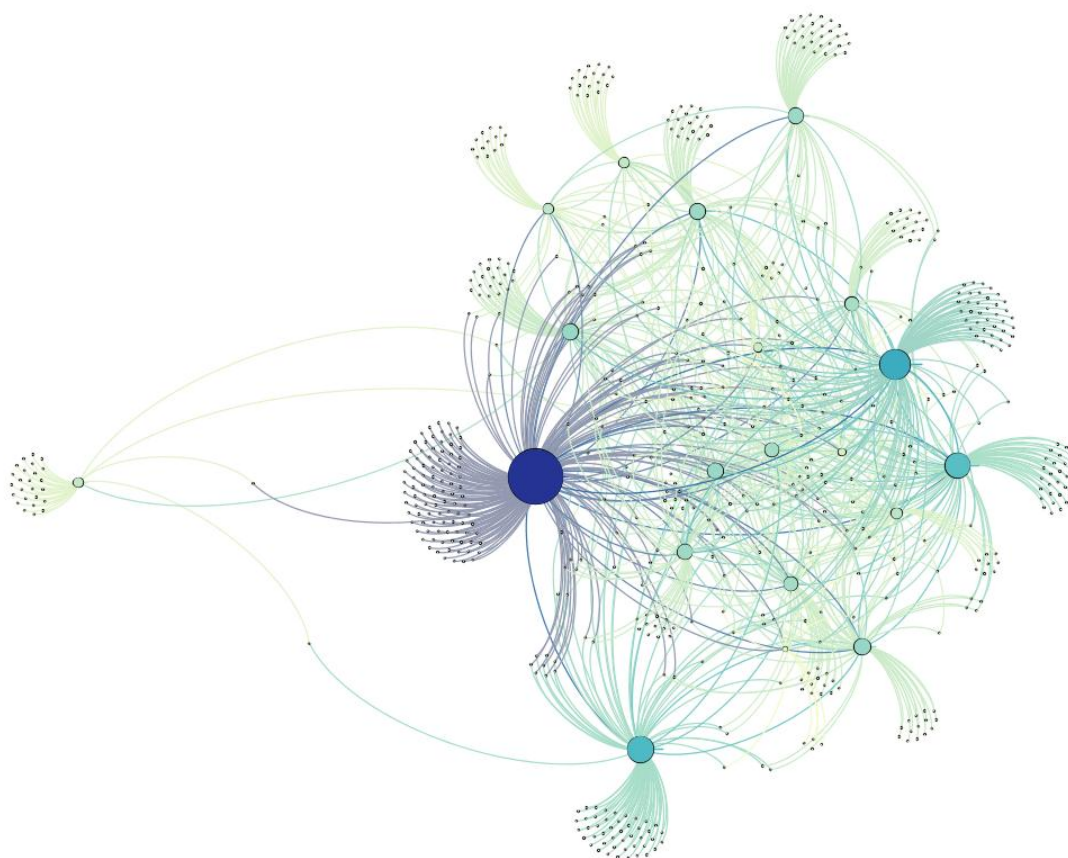


Figure 34. Social diagram considering as “repliers” the top 20 contributors.

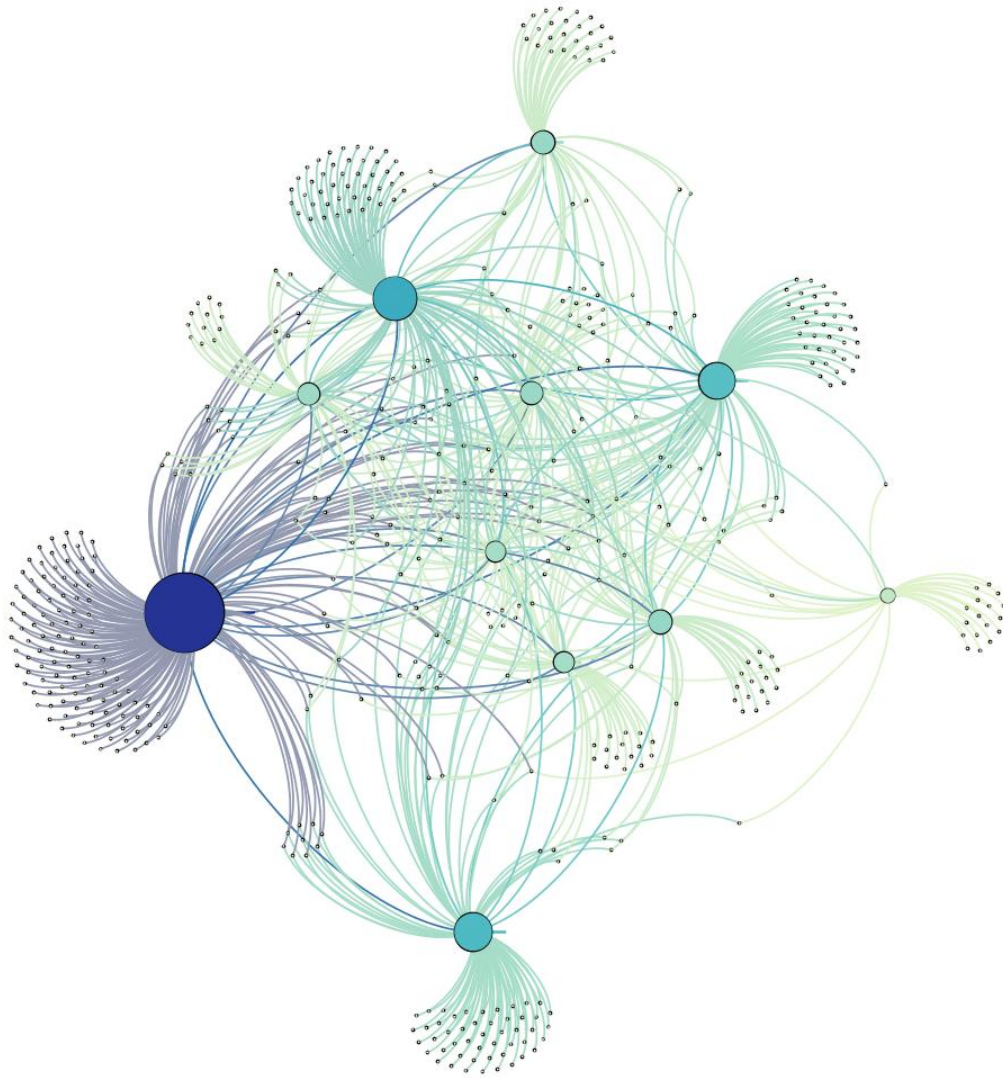


Figure 35. Social diagram considering as “repliers” the top 10 contributors.

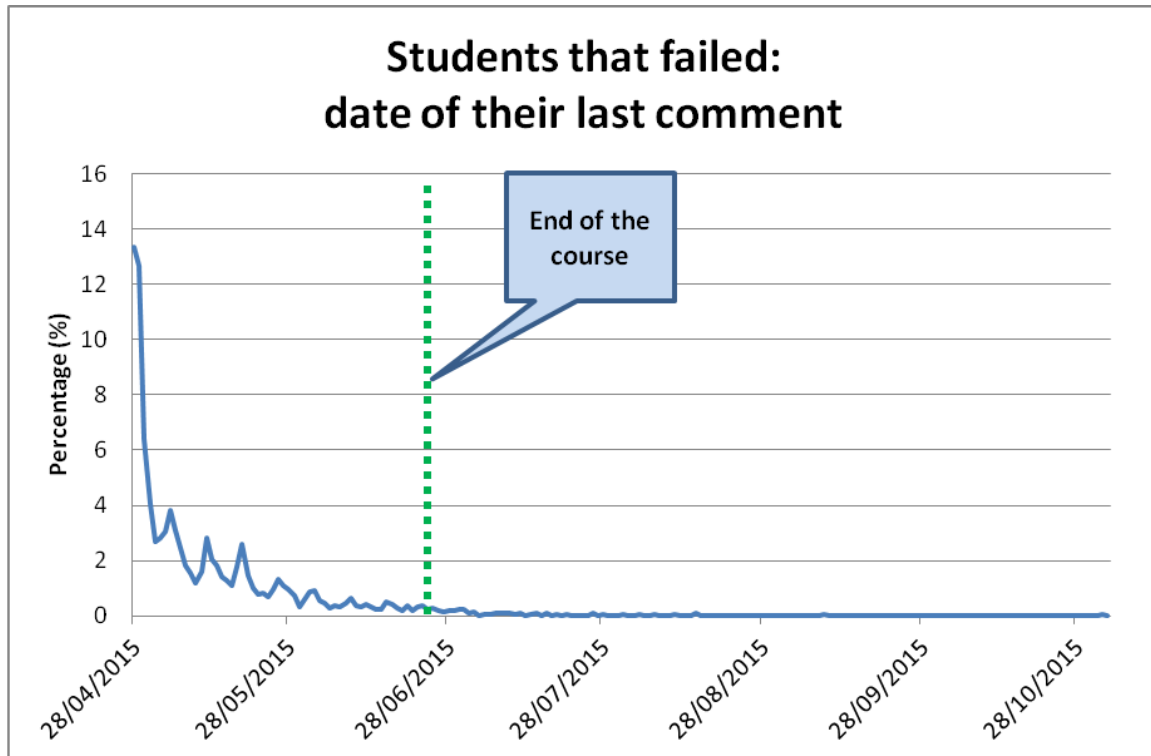


Figure 36. Dates of the last contribution to the forum of the students that failed the course.

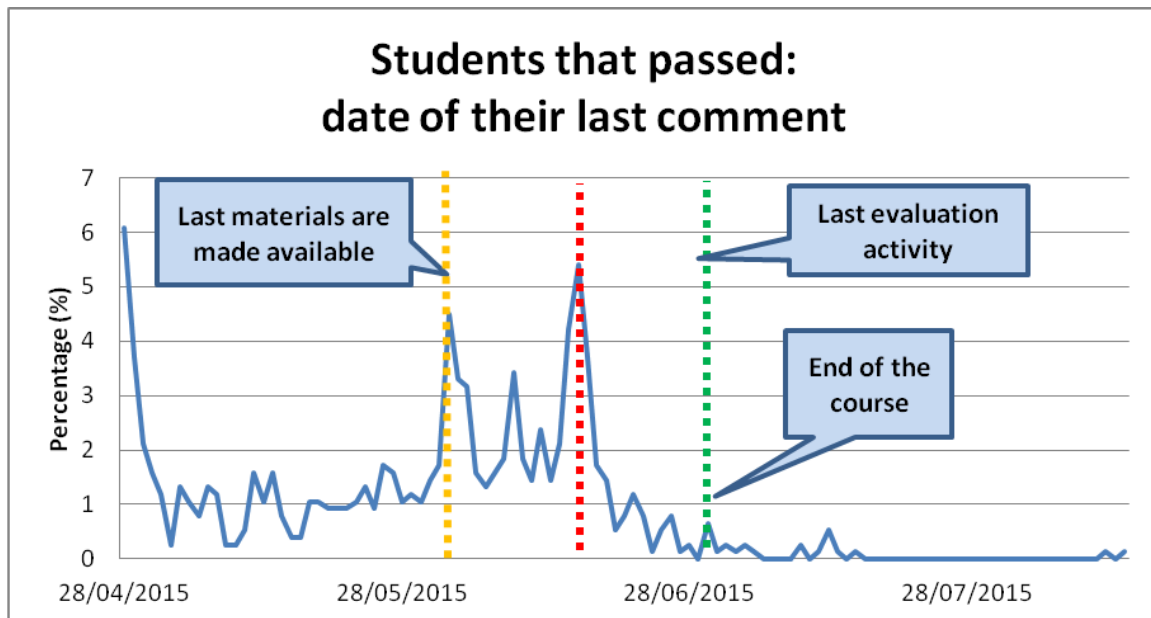


Figure 37. Dates of the last contribution to the forum of the students that passed the course.

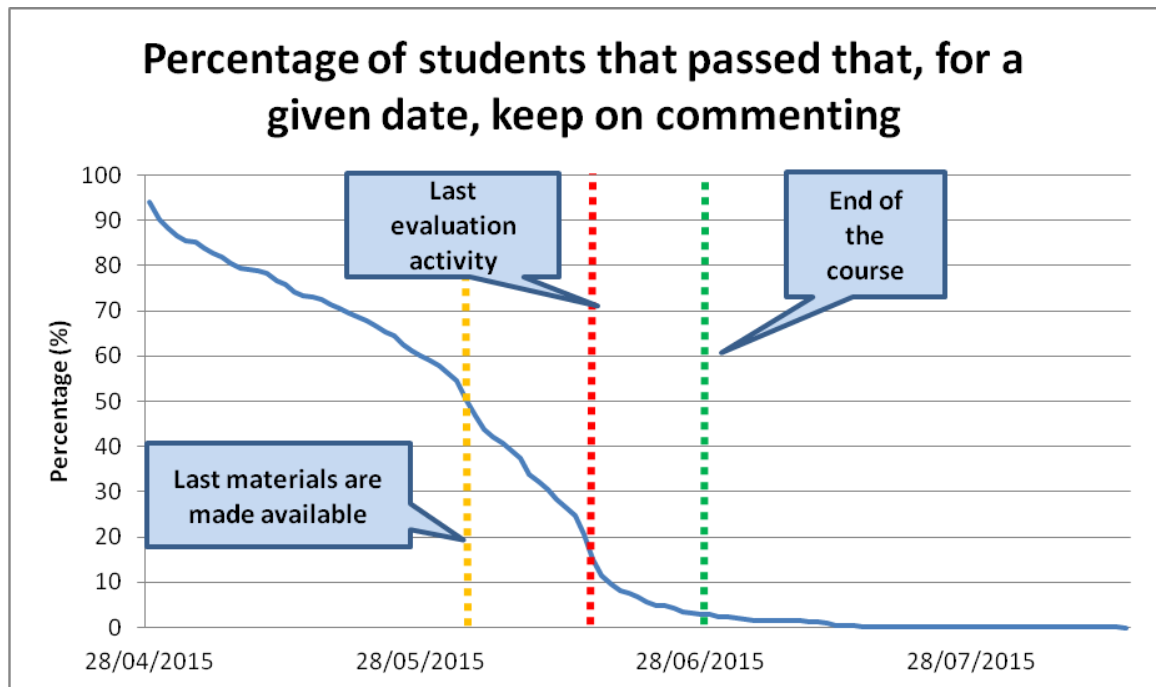


Figure 38. Percentage of students that passed that, for a given date, keep on commenting

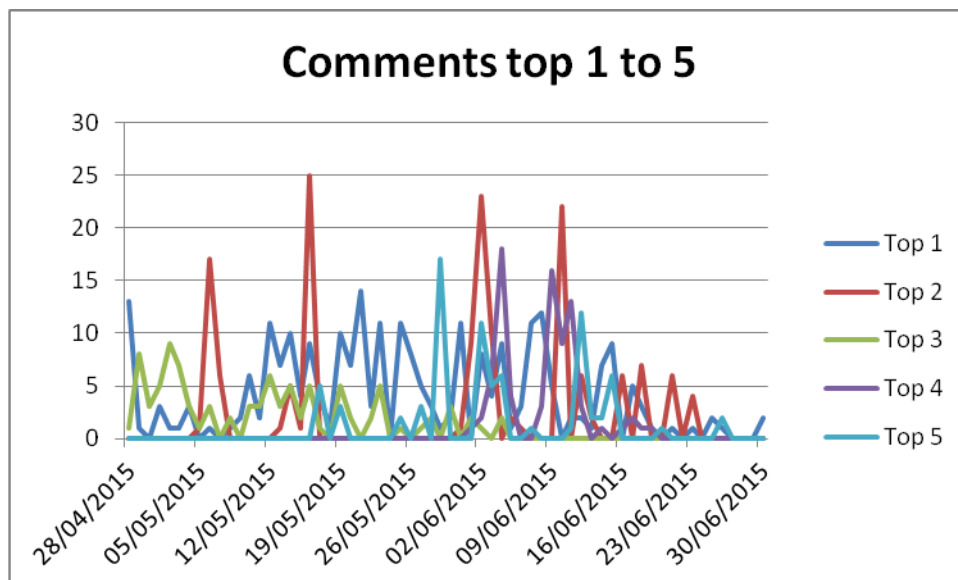


Figure 39. Temporal evolution on the number of comments from top contributor 1 to top contributor 5.

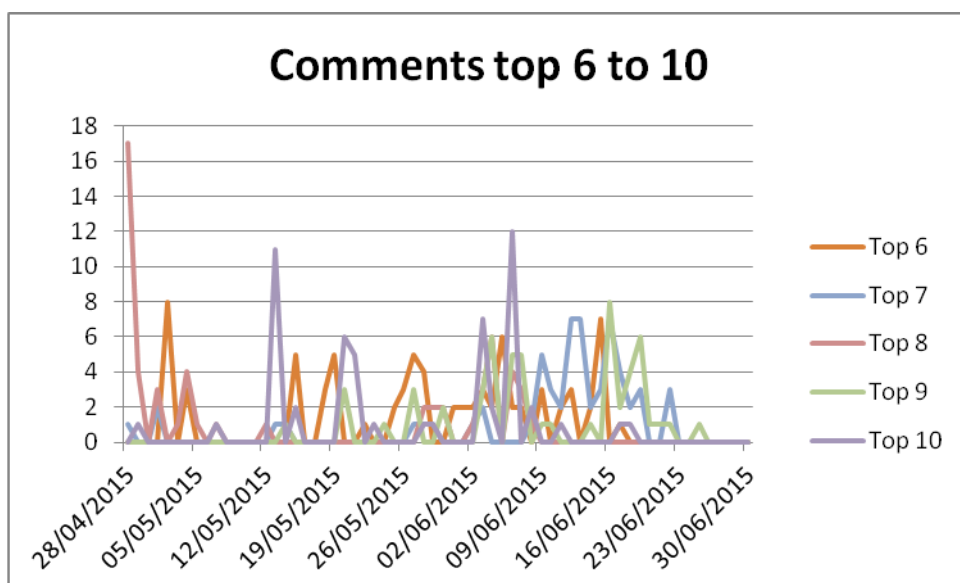


Figure 40. Temporal evolution on the number of comments from top contributor 6 to top contributor 10.

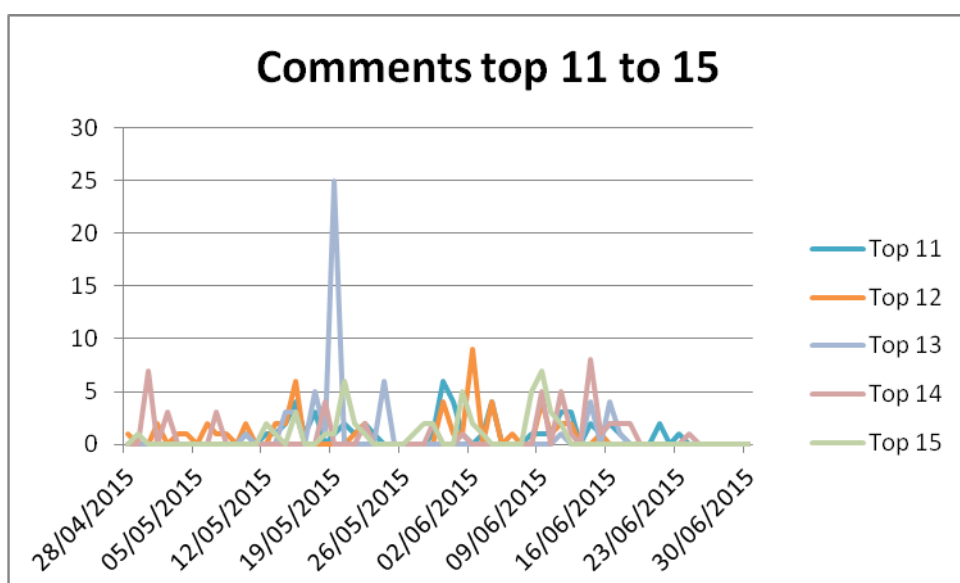


Figure 41. Temporal evolution on the number of comments from top contributor 11 to top contributor 15.

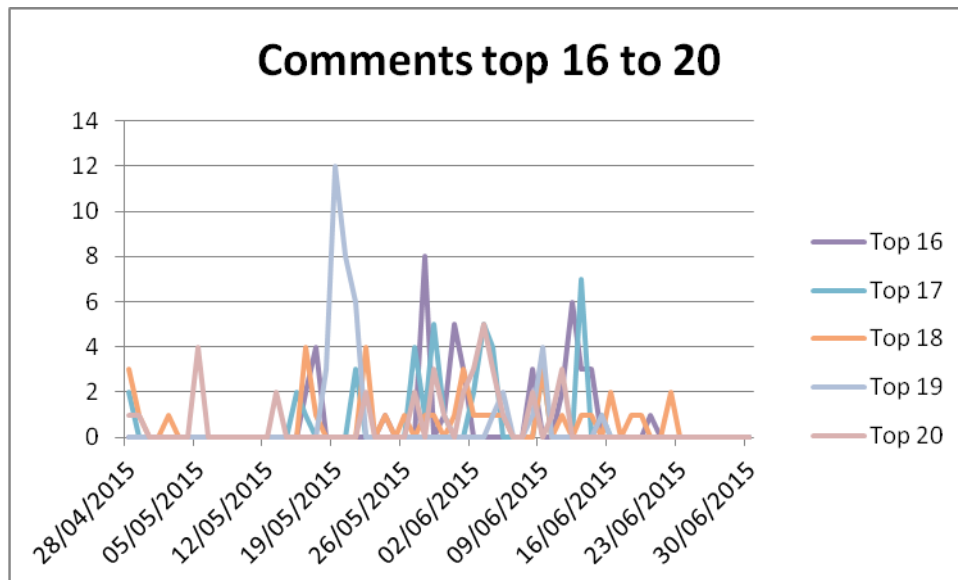


Figure 42. Temporal evolution on the number of comments from top contributor 16 to top contributor 20.

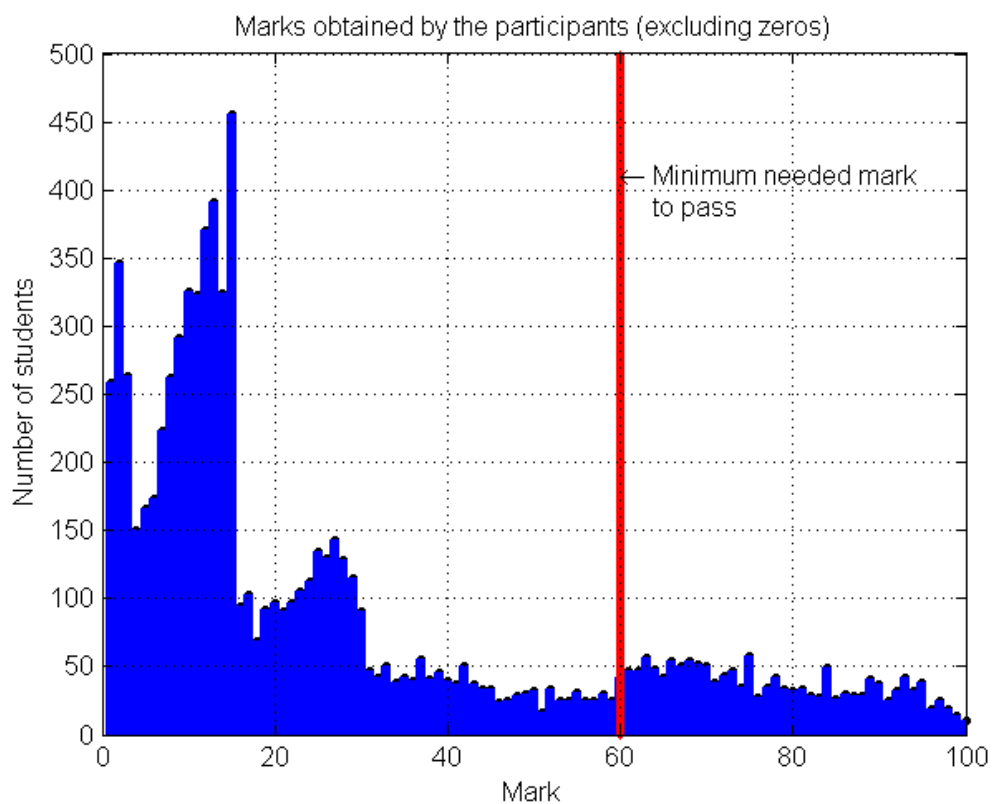


Figure 43. Marks obtained by the students excluding zeros. Note: moderators and staff have not been excluded.

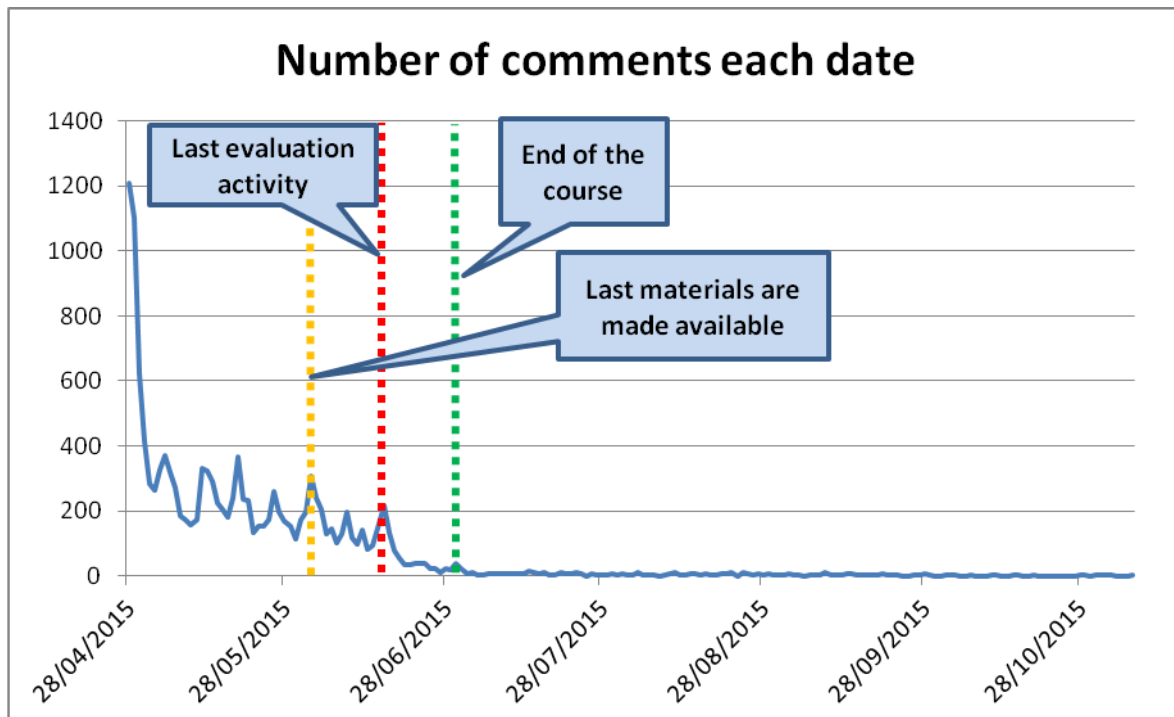


Figure 44. Number of comments each date.

V.5 Results

Firstly, results regarding *Twitter* will be dealt with. There were registered 197 *tweets* from 98 different users, including teachers and staff. The account that wrote the most messages was the account of the course itself (*javaedxuc3m*) with 36 messages, followed by another user (*Tundeiness*, with 13 *tweets*) that happens to be the top contributor number 53. Still, the next user in the list (*hugomesmo*, who wrote 7 *tweets*) happens to be among the top 10 contributors. In Table 11, a classification of these messages has been done.

Type	Number	Percentage
Spread/Greetings	94	47,7
Positives	16	8,1
Negatives	6	3,1
Neutral	15	7,6
Questions/Problems	40	20,3
Answers	26	13,2

Table 11. Classification of the *tweets*.

The most-notable aspect is the large percentage of messages destined to spread the course. Most of these *tweets* were advertisements wrote by the course account itself, others were *retweets* of these and some others were users saying that they had enrolled on the course. With respect to the questions, most of them were related with obtaining a certificate and with asking for some information about evaluation activities.

In fact, it should not be surprising that this social network is more useful as a **communication tool** than as a mechanism for the students to interact with each others. For instance, the restriction of each message only containing at most 140 characters significantly complicates the appearance of “questions and answers” type threads. In fact, the course account replied to a student who had formulated a question suggesting him to write a more-elaborated thread in the forums.

Secondly, **general** data about the course will be dealt with. Excluding teachers and staff, 1522 students passed and 83246 students failed from a total of 84768 users (Figure 45). Among the ones that failed, 76320 of them got a zero as final mark (90% of the total; 91.7% among the participants that failed). Additionally, more information is provided taking into account if the student contributed to the forums or not. There were 79538 users that did not comment. Among them, 764 passed the MOOC (0.96% of 79538). On the other side, among the 5230 users that contributed, 758 passed the course (14.49% of 5230). Besides, another grouping can be done by checking whether the user passed or failed. Among all who passed (1522), 764 did not make any contribution (50.2%) and the remaining 758 did contribute (49.8%; Figure 46). On the other hand, from the participants who failed (83246), 78774 did not write any comment (94.63%) and the remaining 4472 did contribute (5.37%; Figure 47). With respect to users who failed, it is interesting that a large number of them obtained a mark lesser than 16 points (Figure 43), what may mean that most of them took only one evaluation activity (the first one, probably).

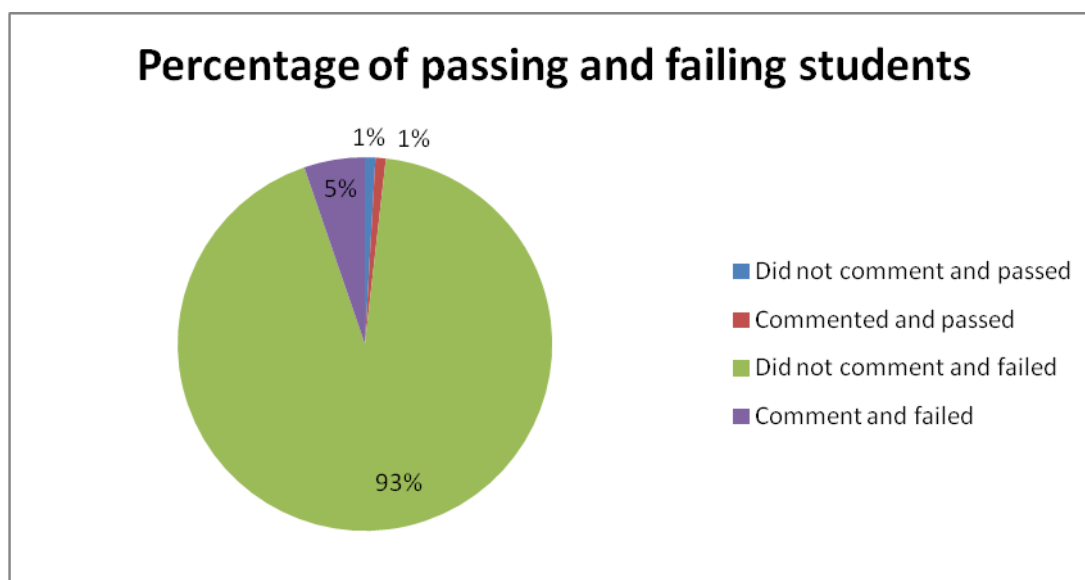


Figure 45. Percentage of passing and failing students.

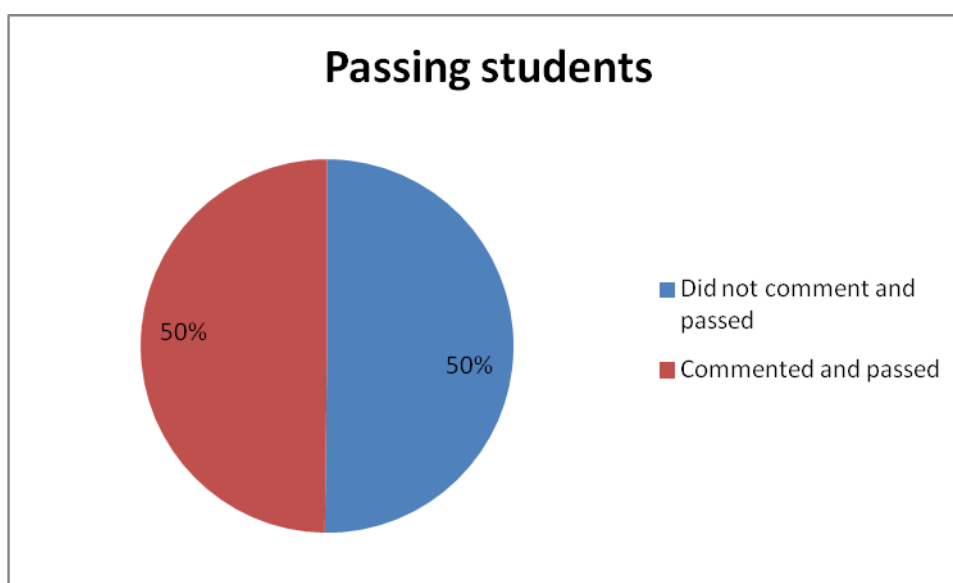


Figure 46. Percentage of passing students that contributed and did not contribute to the forum.

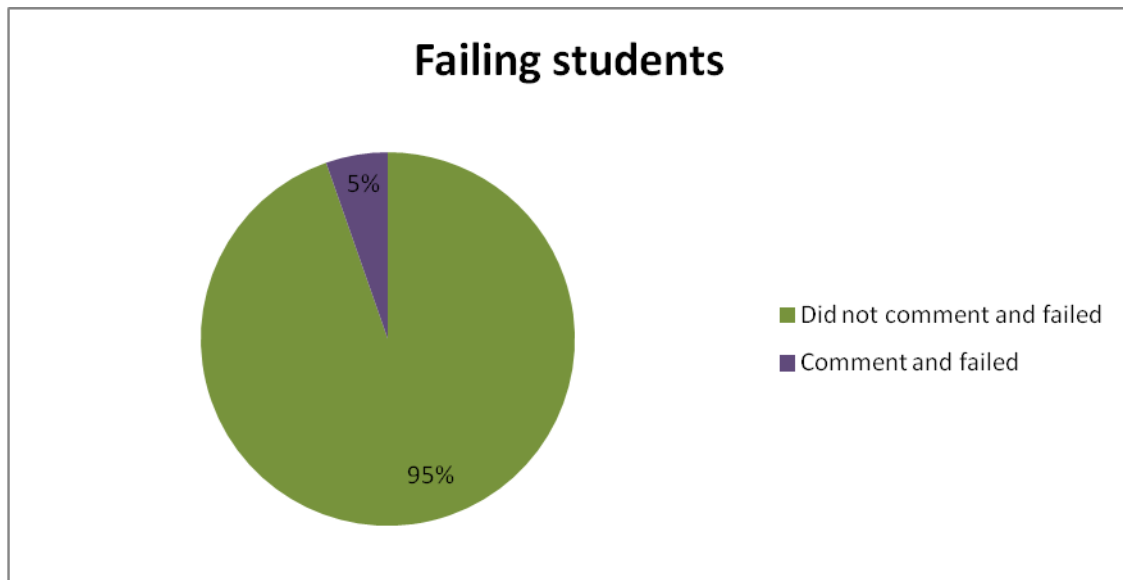


Figure 47. Percentage of failing students that contributed and did not contribute to the forum.

From this point on, the information given will only concern users that made at least one contribution to the forums. Regarding the number of posted messages each day (Figure 44), the peak at the beginning proves the initial hype of the course. From that point on, there are around 200 messages per day. Other peaks usually coincide with days in which new material is provided. Most users only contributed once (Figure 29), which is in agreement with the number of users than comment in the welcoming discussions but then drop the course. As the number of comments increases, the number of users that posted said number of messages decreases, with the larger drop happening between 1 and 2 comments. Respect to comment distributions considering the time in which they were posted (Figure 32), United States' and European's evenings have a greater impact than India's one.

Regarding the participants that passed, it was found that, the more comments a user has posted, the greater the probability that the user passed the course (Figure 27). It was also concluded that the number of users that keeps on contributing decreases as the course draws on (Figure 38). A noticeable drop happened around the date when the last evaluation activity was opened, and, by the time such activity was closed, around 20% of users that passed kept on commenting on the forums.

Finally, some results regarding **top contributors** will be provided. Some of these contribute in a continuous manner while some others comment sporadically. For instance, taking into consideration that the MOOC lasted for 64 days, the user top 1 contributed on 51 of those 64 days while top 4 only commented on 15 days (Figure 39). Actually, the number of days in which contributions were made is lesser than the number of days in which no contributions were made in all top 20 cases except for user top 1 (Table 12). If only the first two weeks of the course are considered, this happens in 4 cases instead of just 1 (Table 13). It is relevant that, during this period, 5 of the 20 "true" top contributors did not make any contribution (maybe they were latecomers), which hinders the aspect of identifying these users early in the course. Another analyzed issue is knowing whether top contributors create new threads or contribute in already created discussions. All top 10 posted more replies than new discussions. The closest numbers are given in user top 8

ANNEX V: SUMMARY

(55.4% answers, 44.6% new discussions), who created many threads to report problems to the teachers and to ask about dates related to the course.

	Contributions	Contributed days	Not contributed days	Contributed days (%)	Not contributed days (%)
Top 1	256	51	13	79,7	20,3
Top 2	155	20	44	31,2	68,8
Top 3	96	29	35	45,3	54,7
Top 4	81	15	49	23,4	76,6
Top 5	79	16	48	25	75
Top 6	78	24	40	37,5	62,5
Top 7	60	22	42	34,4	65,6
Top 8	56	17	47	26,6	73,4
Top 9	56	20	44	31,25	68,75
Top 10	55	16	48	25	75
Top 11	55	29	35	45,3	54,7
Top 12	54	24	40	37,5	62,5
Top 13	54	11	53	17,2	82,8
Top 14	50	16	48	25	75
Top 15	48	19	45	29,7	70,3
Top 16	43	13	51	20,3	79,7
Top 17	42	14	50	21,9	78,1
Top 18	38	24	40	37,5	62,5
Top 19	38	9	55	14,1	85,9
Top 20	36	16	48	25	75

Table 12. Number of contributions to the forum and number and percentage of contributed days of the top 20 contributors. Note: the MOOC lasted for 64 days (from April 28th to June 30th of 2015).

	Contributions	Contributed days	Not contributed days	Contributed days (%)	Not contributed days (%)
Top 1	45	12	3	80	20
Top 2	24	3	12	20	80
Top 3	54	13	2	86,7	13,3
Top 4	0	0	15	0	100
Top 5	0	0	15	0	100
Top 6	11	2	13	13,3	86,7
Top 7	3	2	13	13,3	86,7
Top 8	32	8	7	53,3	46,7
Top 9	0	0	15	0	100
Top 10	2	2	13	13,3	86,7
Top 11	2	2	13	13,3	86,7
Top 12	11	8	7	53,3	46,7
Top 13	1	1	14	6,7	93,3
Top 14	13	3	12	20	80
Top 15	3	2	13	13,3	86,7
Top 16	0	0	15	0	100
Top 17	2	1	14	6,7	93,3
Top 18	5	3	12	20	80
Top 19	0	0	15	0	100
Top 20	6	3	12	20	80

Table 13. Number of contributions to the forum and number and percentage of contributed days of the top 20 contributors during the first two weeks of the MOOC.

It was also studied whether top contributors answered to any user or only to a small group of students. It was found that, even though some participants are only linked to one top contributor, many others are connected to several ones (Figure 34), and there are connections among top contributors themselves, too.

V.6 Conclusions and future work

Before starting with the conclusions, the **primary objectives** will be reminded: on one side, checking whether the hypotheses and conclusions stated by other scientific papers hold for this case of study; on the other side, studying the top contributors' characteristics and trying to predict early whose users will become these more active participants. As the results were organized firstly with general course characteristics and secondly with top contributor's features, the same classification will be followed in this chapter. It is also reminded that the results shown in this section have been obtained by studying a MOOC titled "Introduction to Programming with Java Part 1: Starting to Program with Java".

First of all, the most relevant aspects of the use of the social network *Twitter* will be showed. Clearly, the discussion forum has been more widely used and has had more impact than the “tweets”. In one paper from the related work [13] it was already noted that the use of *Twitter* was lesser than the forum in the MOOC “Digital Education of the Future”. Regarding the topic of these “tweets” (Table 11), 47.7% of these were about communicating and spreading the course. Something similar happened in the just-mentioned course, even though in that case the percentage was lesser since some messages were unrelated to the MOOC, what has not happened in this case. Because of this it is concluded that *Twitter* is more useful as a tool for the course to become known than as a source for links between students to appear; in any case, it is advisable to keep checking these results with the ones from other courses.

As a **general** course feature, only 2% of registered participants passed the MOOC (Figure 45). This fulfils the premise stated in Chapter 2 in which it was said that the number of students that pass a MOOC is typically lower than 10% of the ones that register in it. Making a comparison, in the MOOC “Digital Education of the Future” [13] 8% of registered users passed the course (456 students). Even though these percentages and values are different since the number of enrolled users in each MOOC are far from each other, in both cases less than 10% of participants passed the course, as it was just stated.

In this course, 6% of all participants commented in the forum and the remaining 94% did not (Figure 45). It was computed that 95% of the users that did not comment on the forum failed the course (Figure 47). This goes in accordance with the fact that the vast majority of registered users do not participate in the course (neither in the discussions nor in the activities). Regarding the remaining 5% of students who failed, most of these will probably be users that greeted in the welcoming discussions but then would drop out the course, which appears to be confirmed by Figure 36. On the other side, with respect to the participants that passed, it is interesting that half of them contributed to the forum and the other half did not (Figure 46). This is significantly noteworthy since a highly positive correlation between passing and commenting could be expected; this is, that it existed a greater number of students that passed the course and commented than the number of users that passed and did not comment. Still, if all data is analyzed as a whole (Figure 45), according to the information obtained in this case of study, the following predictions can be done: if a user comments in the forum, there is a 17% chance that the student passes the course and a 83% chance that the user fails; if a participant does not contribute to the forum, there is a 1% chance that he or she passes and a 99% chance that he or she fails. Expressing these conditional probabilities in mathematical form:

$$P(\text{pass}|\text{comment}) = \frac{1}{2} = 50\% \qquad P(\text{comment}|\text{pass}) = \frac{1}{6} \approx 17\%$$

With respect to learners’ final marks (Figure 43), as it was mentioned in Chapter 5, it can be deduced that most of participants that do not obtain a zero as final mark only take one evaluation activity, probably the first one, before disengaging from the course. Also, the correlation between final mark and number of contributions was computed, obtaining the value 0.21. Even though it is a positive value, it is also low, and therefore further statistical analyses are required in order to obtain meaningful conclusions. The same happens when using a relation between created threads and comments published in already created discussion, whose correlation with final marks is 0.25.

Continuing with conclusions, the graph that represents the evolution of the number of comments according to the passing of time (Figure 44) corresponds to what usually happens in these MOOCs (again, due to the excitement produced at the beginning of the course and the welcoming discussions). In this line, there is a large portion of the messages during the first days and then a pronounced drop takes place. Later, some peaks are reached, especially around days in which new material is provided. In agreement with this aspect and with the previous passing and failing classification, it is known that most participants (3699, 68% of all) only contributed once (Figure 29). The number of users that commented twice is reduced to 689 (13% of all). The mean is 2.5 messages per user (taking only into account the students that contributed).

With respect to the time periods in which the number of messages is greater, it should obviously be taken into account the time difference among distinct zones of the world (Figure 32). In this study it was found that United States and Europe's evenings have more weight than India's one (these locations have been selected as references since most of the users belong to these places).

As stated before, that data from this case of study indicates that 6% of registered users made at least one contribution to the forum, even though only 1% of all commented and passed the MOOC. In line with this aspect, it was calculated that, the more contributions a user has made, the greater the probability that this user has passed (Figure 27). Intuitively, this corresponds with the students' engagement in the MOOC. For instance, in participants that made more than 9 contributions the uncertainty of passing or failing is at 50%. From 3 comments on, the graph can be approximated as a line (until reaching 100% of passing users). Something similar happens if only the first two weeks of the course are taken into account, although the approximation to a line is less reliable. Regarding data from the entire course, it can be easily obtained the equation of the aforementioned line (reminding that it should not be applied for percentages lesser than 30%):

$$\left. \begin{array}{l} P(\text{pass} | \text{comment} > 9 \text{ times}) = 50\% \\ P(\text{pass} | \text{comentar} > 24 \text{ times}) = 75\% \\ P(\text{pass} | \text{comment} > 36 \text{ times}) = 100\% \end{array} \right\} \rightarrow \% \text{ Passing} \approx 1,9 \cdot \text{AmountComments} + 32,9$$

With respect to the students that both passed the course and commented in the forum, data relating their last comment date was computed (Figure 37). On the last day available to take the last evaluation activity, around 20% of users that commented and passed kept contributing (Figure 38). It could be underlined that, even though the number of participants contributing decreases as the course goes by, there exists a notable drop around the 2nd of June, last date in which new material was provided. Thus, it is possible that some users have taken the evaluations activity related to this material shortly after it was available and then they have not checked the forums anymore.

Before providing specific data about top contributors, it will be reminded that the early identification of such contributors is related to awarding them with special moderator permissions in the discussion forums (the platform *edX* calls them "Community Teaching Assistant" or "Community T.A."). Particularizing obtained information on **top contributors**, 17 of them contributed at least 25% of the days the

course lasted. Some of them commented more frequently than others and some of them contributed sporadically. In addition, the early identification of these users is a complex task according to the obtained data. For instance, 5 of 20 top contributors did not comment during the first two weeks of the course (maybe they were latecomers, what would strengthen the importance of this group as stated in some papers [13]), and another 6 of them made less than 6 contributions (Table 13). If these top contributors are sorted in descendent number of comments taking into account only the first two, among them can be found 8 of the “true” top contributors of all MOOC (specifically, tops 1, 2, 3, 6, 8, 12, 13 and 20). This way, if early in the course the top 10 were selected, among them there would be 5 of the “true” top contributors.

Regarding their contributions to the forum, the top 10 contributors have posted messages in already created discussions more than they have create new threads (Table 14); only in 3 of them the percentage of responses is lesser than 75% and the percentage of started discussions greater than 25%. In one case, the percentage of started threads is that of 44.6% of total published messages; this user usually created threads in which he reported errors to the staff or asked information about the publication dates of some material.

	Answers in threads	Starts of new discussions	Answers in threads (%)	Starts of new discussions (%)
Top 1	238	18	93	7
Top 2	124	31	80	20
Top 3	92	4	95,8	4,2
Top 4	77	4	95,1	4,9
Top 5	75	4	94,9	5,1
Top 6	52	26	66,7	33,3
Top 7	46	14	76,7	23,3
Top 8	31	25	55,4	44,6
Top 9	50	6	89,3	10,7
Top 10	41	14	74,5	25,5

Table 14. Number and percentage of started discussions and comments in already created discussions of the top 10 contributors.

With respect to links formed by the top contributors with other users (Figure 34), it can be concluded that these contributors do not seem to focus in just a few users. Thus, even though there are participants that are only connected to one top contributors, many of them have connections with more than one. Even there exists links among the top contributors themselves.

Finally, as future work, the main line that can be followed is the **study of top contributors** in different ways. Inside this very same MOOC, additional studies can be done, like inspecting how technical the contributions are. In this line, it could be checked if each one comments only about one issue or not. Another aspect that was not taken into account in this work was top contributors’ profiles. Thus, if more personal information about these contributors was studied, there could be some common features among them (above all, it may be relevant taking into consideration their previous training or studies).

In more technical lines and aspects whose automation would be simple, new algorithms could be programmed. It may be interesting to know the time periods in which top contributors comment (that is, obtaining a graph similar to Figure 32 taking into account only top contributors). The contributions that have more positive votes could be selected and then check to which contributors they belong. In addition, along this line, the field “endorsed” could also be taken into account for a similar analysis (it is reminded that, when this field is followed by “true”, a moderator or teacher has marked a message as a valuable contribution). With respect to more complex statistical studies, more correlation analyses could be made by selecting only the number of contributions and the marks of each week; that is, checking the marks of the activities of each week and relating them to the number of messages associated to that week.

Once several studies about identifying top contributors have been carried out, and the amount of information is large enough to make a train group and a test group, analyses like linear regressions and descriptors extraction could be made. In this sense, an algorithm that adequately classifies the train group and then evaluates the test group would be programmed. If no satisfactory results are obtained, the code is modified and the program is executed again. It could be interesting to have a very high detection probability in exchange of having a quite high false alarm probability (that is, detecting practically all top contributors but selecting also some users that do not belong to this group). Besides, the results could be refined by making similar analyses each week and by considering all accumulated course data.

Outside this MOOC, it could be analyzed how the users categorized here as top contributors behave in other courses. Also, it could be checked if these users actively use social networks like *Twitter* in their daily life. Of course, it could be studied whether the data obtained in this project resemble other courses or not. It would be interesting to divide these courses between the ones with topics that are similar to the one this work studies and courses with different subjects (this is, it may happen that the interaction among users changes drastically in a MOOC about Programming and a MOOC about Medicine).